Probabilistic Models in Weather Forecasting

Dissertation submitted for the degree of

"Doctor of the Hungarian Academy of Sciences"

Sándor Baran

Debrecen, 2019

dc_1665_19

Acknowledgements

Here I would like to thank all the people who have, either directly or indirectly, contributed to this dissertation.

First of all to my wife, Ágnes for her love, patience, understanding and standing always by me.

To my daughters Zsuzsanna and Júlia for bringing so much joy into my life.

To my parents for helping and encouraging me.

To Tilmann Gneiting, who motivated my research in probabilistic forecasting and supervised my first steps in this area, hosted me several times in Heidelberg, and gave an enormous help with his comments, remarks and suggestions.

To András Horányi for introducing me atmospheric science, his help and collaboration.

To Sebastian Lerch for the fruitful collaboration, great ideas and all his help with the R codes.

To Annette Möller for her help with copula models and for collaboration.

To Stephan Hemri for introducing me hydrological forecasting and helping with the analog-based approaches.

To Martin Leutbecher and Zied Ben Bouallègue for the collaboration and for hosting me several times at the European Centre for Medium-Range Weather Forecasts.

To all current and former members of the Computational Statistics Group of the Heidelberg Institute for Theoretical Studies for providing an inspiring working environment. Besides my collaborators special thanks to Kira Feldmann, Alexander Jordan, Fabian Krüger and Roman Schefzik.

To my current and former students Mailiu Díaz Peña, Mehrez El Ayari, Dóra Nemoda and Marianna Szabó for their hard work in the field of probabilistic forecasting and tolerating me as their supervisor.

Further, I would like to thank the following organizations and projects for providing the financial background of my research.

- DAAD programs "Research Stays for University Academics and Scientists, 2015" and "Research Stays for University Academics and Scientists, 2017."
- EFOP-3.6.1-16-2016-00022 project co-financed by the European Union and the European Social Fund.
- EFOP-3.6.3-VEKOP-16-2017-00002 project co-financed by the Hungarian Government and the European Social Fund.
- European Centre for Medium-Range Weather Forecasts.
- Hungarian Scientific Research Fund under Grant No. OTKA NK101680.
- János Bolyai Research Scholarship of the Hungarian Academy of Sciences (2015–2018).
- National Research, Development and Innovation Office under Grant No. NN125679.
- Visiting Scientist Program of the Heidelberg Institute for Theoretical Studies for years 2016 and 2018.
- TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project supported by the European Union and co-financed by the European Social Fund.

Finally, I am grateful

to the University of Washington MURI group for providing the University of Washington mesoscale ensemble data;

to Mihály Szűcs from the Hungarian Meteorological Service for the ALADIN-HUNEPS data;

to the German Federal Office of Hydrology, and in particular Bastian Klein, for providing the hydrological data.

List of abbreviations

ALADIN-HUNEPS	Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System
ALARO	Aire Limitée Application de la Recherche à l'Operationnel
ARPEGE	Action de Recherche Petite Echelle Grande Echelle
BfG	German Federal Institute of Hydrology
BFGS	Broyden-Fletcher-Goldfarb-Shanno (algorithm)
BMA	Bayesian Model Averaging
BS	Brier Score
BSS	Brier Skill Score
CDF	Cumulative Distribution Function
COSMO	Consortium for Small-Scale Modelling
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
CSG	Censored and Shifted Gamma (distribution)
DM	Diebold-Mariano (test)
DS	Determinant Sharpness
DST	Daylight Saving Time
ECMWF	European Centre for Medium-Range Weather Forecasts
EE	Euclidean Error
EM	Expectation-Maximization (algorithm)
EMOS	Ensemble Model Output Statistics
ENS	(51-member ECMWF) Ensemble
EPS	Ensemble Prediction System
ES	Energy Score
GEFS	Global Ensemble Forecast System
GEV	Generalized Extreme Value (distribution)
GLAMEPS	Grand Limited Area Model Ensemble Prediction System
HIRLAM	High Resolution Limited Area Modelling
HMS	Hungarian Meteorological Service
HRES	(ECMWF) High-Resolution (forecast)
ISBA	Intéractions Soil Biosphere Atmosphère (parametrization scheme)
LEPS	Limited-Area Ensemble Prediction System
LN	Log-Normal (distribution)
LogS	Logarithmic Score

MAE	Mean Absolute Error
ML	Maximum Likelihood
NCEP	National Center for Environmental Prediction
NWP	Numerical Weather Prediction
PDF	Probability Density Function
PEARP	Prévision d'Ensemble ARPEGE
PIT	Probability Integral Transform
RMSE	Root Mean Square Error
STRACO	Soft Transition Condensation (parametrization scheme)
SURFEX	Surface Externalisée (parametrization scheme)
TN	Truncated Normal (distribution)
twCRPS	Threshold-Weighted Continuous Ranked Probability Score
twCRPSS	Threshold-Weighted Continuous Ranked Probability Skill Score
UTC	Coordinated Universal Time
UWME	University of Washington Mesoscale Ensemble

Contents

Introduction

Capturing and modelling uncertainty is an essential need in any forecasting problem, and in weather or hydrological prediction it may result in an enormous economical benefit. In the early 90's there was an important shift in the practice of weather forecasting from deterministic forecasts obtained using numerical weather prediction (NWP) models in the direction of probabilistic forecasting. The crucial step was the introduction of ensemble prediction systems (EPSs) in operational use in 1992 both at the European Centre for Medium-Range Weather Forecasts (ECMWF) and the U.S. National Meteorological Center. An EPS provides a range of forecasts corresponding to different runs of the NWP models, which are usually generated from random perturbations in the initial conditions and the stochastic physics parametrization. In the last decades, the ensemble method has become a widely used technique all over the world as using ensemble forecasts one can, for instance, easily provide prediction intervals reflecting to forecasts uncertainty. The advantage of ensemble forecasting is nicely illustrated by Figure 1 showing point forecasts of temperature for Debrecen, Hungary, and the corresponding ECMWF ensemble forecasts. In the latter case (Figure 1b) the ensemble mean can serve as a point forecast, however, one can also observe how the forecast uncertainty increases with the increase of the lead time of prediction. Note that both forecasts are available for public on the official web page of the Hungarian Meteorological Service (`www.met.hu`).

However, raw ensemble forecasts often exhibit systematic errors as they might be biased or badly calibrated calling for some form of post-processing. Simple approaches to bias correction or calibration have a long history, however, in the first years of the XXI. century several more sophisticated methods appeared (Wilks, 2006), including parametric models providing full predictive distributions of the weather variables at hand. Starting with the fundamental works of Tilmann Gneiting and Adrian Raftery (Gneiting and Raftery, 2005; Gneiting *et al.*, 2005; Raftery *et al.*, 2005) introducing Bayesian model averaging (BMA) and ensemble model output statistics (EMOS) for ensemble calibration, statistical post-processing of ensemble forecasts became a hot topic both in statistics and in atmospheric sciences, resulting in a multitude of probabilistic models for different weather quantities, new methods of estimation of parameters of these models and novel approaches to forecast verification (Buizza, 2018). Recently, the German Meteorological Service uses a special EMOS post-processing model (Schuhen *et al.*, 2012) in operational wind vector prediction for the Frankfurt Airport, and there are ongoing research projects e.g. at the ECMWF pointing towards the introduction of statistical calibration in operational use (see e.g. Gneiting, 2014; Richardson *et al.*, 2015; Baran *et al.*, 2019b).

This dissertation is a summary of the achievements of the author in the area of probabilistic forecasting. It contains new BMA and EMOS models for post-processing of water levels and different weather quantities, novel approaches to training data selection in the

(a) (b)

Figure 1: Point forecasts (a) and ECMWF ensemble forecasts (b) of temperature for Debrecen, Hungary. On panel (b) light and dark orange belts correspond to 80 % and 50 % central prediction intervals. Source: `www.met.hu`

parameter estimation process, and in some cases efficient algorithms for parameter estimation are also provided. Note that the presented results are purely applied. Due to the nature of the problems investigated, the verification of the proposed models and algorithms can be based only on carefully chosen case studies, which is a standard approach in probabilistic weather prediction. Using appropriate verification scores the forecast skill of each suggested method is compared with the predictive performance of the corresponding state of the art calibration models. The current work is mainly based on eight papers of the author published either in statistical journals (*Computational Statistics and Data Analysis, Environmetrics, Journal of the Royal Statistical Society: Series C*) or in journals in the field of atmospheric (*Meteorology and Atmospheric Physics, Quarterly Journal of the Royal Statistical Society*) or water sciences (*Water Resources Research*), but some results of other published journal articles are also used.

The dissertation consists of six main chapters. Chapter 1 introduces the basic notions of ensemble forecasting and statistical post-processing, lists the main parametric post-processing approaches and parameter estimation strategies and describes the methods of forecast evaluation. In Chapter 2 a novel BMA post-processing model for calibration of ensemble forecasts of water levels is proposed (Baran *et al.*, 2019a) together with an efficient expectation-maximization (EM) algorithm based maximum likelihood (ML) method for parameter estimation. Chapter 3 deals with calibration of wind speed forecasts. It describes a new BMA approach (Baran, 2014) and two different EMOS models (Baran and Lerch, 2015, 2016) together with the existing up to date methods. In Chapter 4 a novel EMOS model for probabilistic quantitative precipitation forecasting (Baran and Nemoda, 2016) is compared with the existing parametric approaches. Joint calibration of wind speed and temperature ensemble forecasts is investigated in Chapter 5 by introducing bivariate BMA (Baran and Möller, 2015) and EMOS (Baran and Möller, 2017) models and comparing their predictive performance with the more general Gaussian copula approach (Möller *et al.*, 2013). The fundamental part of the dissertation ends with Chapter 6, where two semi-local methods for choosing training data for post-processing models are described, followed by a short chapter containing some general conclusions.

Finally, we would like to mention that the implementation of the presented methods resulted in thousands of lines of `R` code (R Core Team, 2019) and most of the EMOS approaches considered in Chapters 3 and 4 are now available to a wide range of users as parts of the `ensembleMOS` package (Yuen *et al.*, 2018) of `R`.

Chapter 1

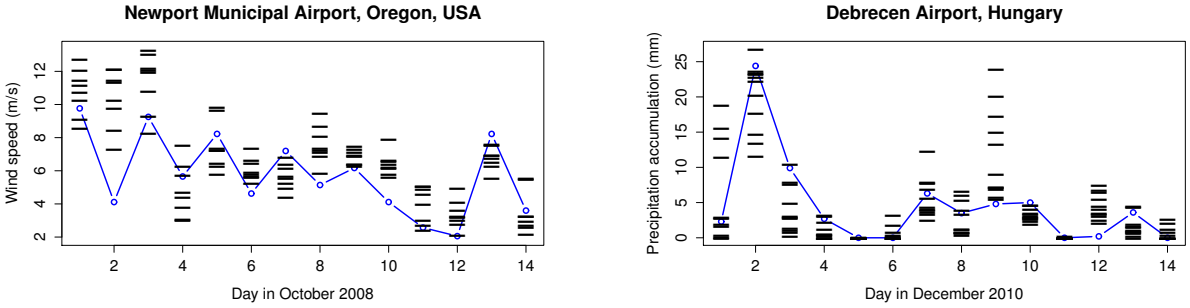
Probabilistic forecasts and forecast evaluation

1.1 Ensemble forecasts

The main objective of weather forecasting is to give a reliable prediction of future atmospheric states on the basis of observational data, prior forecasts valid for the initial time of the predictions, and mathematical models describing the dynamical and physical behaviour of the atmosphere. These models numerically solve the set of the hydrothermodynamic non-linear partial differential equations of the atmosphere and its coupled systems. A disadvantage of these NWP models is that since the atmosphere has a chaotic character the solutions depend on the initial conditions and also on other uncertainties related to the numerical weather prediction process. In practice it means that the results of such models are never fully accurate and the forecast uncertainties should be also taken into account in the forecast preparation. One can reduce the uncertainties by running the model with different initial conditions resulting in an ensemble of forecasts (Leith, 1974). Using a forecast ensemble one can estimate the probability distribution of future weather variables which opens the door for probabilistic weather forecasting (Gneiting and Raftery, 2005), where not only the future atmospheric states are predicted, but also the related uncertainty information such as variance, probabilities of various events, etc.

Since its first operational implementation (Buizza *et al.*, 1993; Toth and Kalnay, 1997), this approach has became a routinely used technique all over the world and recently all major weather prediction centres have their own operational EPSs, e.g. the 30-member Consortium for Small-scale Modelling (COSMO-DE) EPS of the German Meteorological Service (Gebhardt *et al.*, 2011; Ben Bouallègue *et al.*, 2013), the 35-member Prévision d'Ensemble ARPEGE[1] (PEARP) EPS of Méteo France (Descamps *et al.*, 2015) or the 51-member EPS of the independent intergovernmental ECMWF (ECMWF Directorate, 2012; Molteni *et al.*, 1996; Leutbecher and Palmer, 2008), whereas the Hungarian Meteorological Service (HMS) operates the 11-member Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS; Horányi *et al.*, 2006). It is also worth mentioning the experimental 8-member University of Washington mesoscale ensemble (UWME; Eckel and Mass, 2005), as an example of an EPS operated not by a weather centre.

[1] Action de Recherche Petite Echelle Grande Echelle (i.e. Research Project on Small and Large Scales)



(a) (b)

Figure 1.1: (a) Wind speed observations (blue line) and corresponding UWME forecasts (bars) for Newport Municipal Airport, Oregon, USA, for the first two weeks of October 2008; (b) observed precipitation accumulation (blue line) and the corresponding ALADIN-HUNEPS ensemble forecasts (bars) for Debrecen Airport, Hungary, for the first two weeks of December 2010.

Although the transition from single deterministic forecasts to ensemble predictions can be seen as an important step towards probabilistic forecasting, ensemble forecasts are often underdispersive, that is, the spread of the ensemble is too small to account for the full uncertainty, and subject to systematic bias. This phenomenon has been observed with several operational ensemble prediction systems (see e.g. Buizza *et al.*, 2005; Park *et al.*, 2008; Bougeault *et al.*, 2010). A possible solution to account for this deficiency is some form of statistical post-processing (Buizza, 2018).

To illustrate the systematic errors of ensemble forecasts, Figure 1.1a shows UWME wind speed forecasts for Newport Municipal Airport (OR) and the corresponding observations for the first two weeks of October 2008, and Figure 1.1b shows ALADIN-HUNEPS forecasts of precipitation accumulation at Debrecen Airport and the corresponding observations for the first two weeks of December 2010. Both time series illustrate the lack of an appropriate representation of the forecast uncertainty as the verifying observations frequently fall outside the range of the ensemble forecasts.

1.2 Statistical post-processing

Over the past decade, various statistical post-processing methods have been proposed in the meteorological and statistical literature, for an overview see e.g. Wilks (2006); Gneiting (2014); Williams *et al.* (2014), or Vannitsem *et al.* (2018). Among these probably the most popular parametric approaches are the BMA (Raftery *et al.*, 2005) and the EMOS or non-homogeneous regression (Gneiting *et al.*, 2005), which are partially implemented in the `ensembleBMA` (Fraley *et al.*, 2011) and `ensembleMOS` (Yuen *et al.*, 2018) packages of `R` (R Core Team, 2019) and provide estimates of the probability distributions of the predictable weather quantities. Once the predictive distribution is given, its functionals (e.g. median or mean) can easily be calculated and considered as point forecasts.

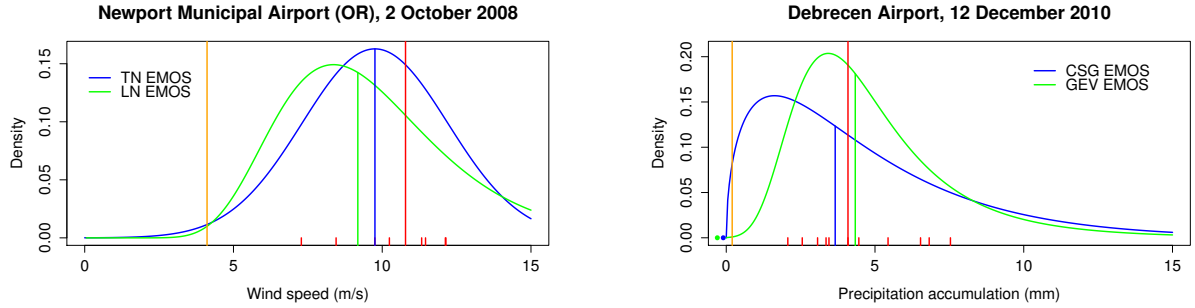
The BMA predictive probability density function (PDF) of a future weather quantity is the weighted sum of individual PDFs corresponding to the ensemble members. An individual PDF can be interpreted as the conditional PDF of the future weather quantity

provided the considered forecast is the best one and the weights are based on the relative performance of the ensemble members during a given training period. In this way BMA is a special, fixed parameter version of dynamic model averaging method developed by Raftery *et al.* (2010). Weights and other model parameters are usually estimated using linear regression and ML method, where the maximum of the likelihood function is found by the EM algorithm. We remark that due to their flexibility, mixture models play an essential role in data analysis (Böhning, 2014) and parameter estimation in mixture models is a typical application of the EM algorithm (see Dempster *et al.* (1977), McLachlan and Krishnan (1997) or more recently Lee and Scott (2012), Chen and Lindsay (2014)). The BMA models of various weather quantities differ only in the PDFs of the mixture components. For temperature and sea level pressure a normal distribution provides an appropriate model (Raftery *et al.*, 2005), but different laws are needed for wind speed (Sloughter *et al.*, 2010; Baran, 2014), precipitation (Sloughter *et al.*, 2007) or surface wind direction (Bao *et al.*, 2010). However, one should also mention that in some situations BMA post-processing might result, for instance, in model overfitting (Hamill, 2007) or overweighting climatology (Hodyss *et al.*, 2016).

The essentially simpler EMOS approach uses a single parametric distribution as a predictive PDF with parameters depending on the ensemble members. The unknown parameters specifying this dependence are estimated using forecasts and validating observations from a rolling training period, which allows automatic adjustments of the statistical model to any changes of the EPS (for instance seasonal variations or EPS model updates). Similar to the BMA approach, different weather quantities require different predictive PDFs. For example, Gneiting *et al.* (2005) models temperature with a Gaussian predictive distribution where the mean is an affine function of the ensemble member forecasts and the variance is an affine function of the ensemble variance. Over the last years the EMOS approach has been extended to other weather variables such as wind speed (Thorarinsdottir and Gneiting, 2010; Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2015; Scheuerer and Möller, 2015), precipitation (Scheuerer, 2014; Scheuerer and Hamill, 2015; Baran and Nemoda, 2016), and total cloud cover (Hemri *et al.*, 2016).

To illustrate the EMOS approach to post-processing, Figure 1.2a shows the observed wind speed, the corresponding UWME forecasts and truncated normal (TN) and log-normal (LN) EMOS predictive distributions (for details see Sections 3.2.1 and 3.2.2, respectively) for Newport Municipal Airport for 2 October 2008. A different situation is shown in Figure 1.2b, where the observed precipitation accumulation, the corresponding ALADIN-HUNEPS ensemble forecasts and estimated censored and shifted gamma (CSG) and censored generalized extreme value (GEV) EMOS predictive distributions (see Sections 4.2.1 and 4.2.2, respectively) for Debrecen Airport for 12 December 2010 are plotted. In both cases, the spread of the ensemble forecasts is notably smaller than the spread of the post-processed forecast distribution. Note that in both examples two different EMOS models are proposed for the same weather quantity and in general, the success of statistical post-processing relies on finding appropriate parametric families for the weather variable of interest. However, the choice of a suitable parametric model is a non-trivial task and often a multitude of competing models is available. The relative performances of these models usually vary for different data sets and applications.

The regime-switching combination models proposed by Lerch and Thorarinsdottir (2013) and also investigated by Baran and Lerch (2015) partly alleviate the limited flexibility of single parametric family models by selecting one of several candidate models



(a) (b)

Figure 1.2: (a) Wind speed observations, the corresponding UWME forecasts and TN and LN EMOS predictive distributions for Newport Municipal Airport (OR) for 2 October 2008; (b) observed precipitation accumulation, the corresponding ALADIN-HUNEPS ensemble forecasts and CSG and GEV EMOS predictive distributions for Debrecen Airport for 12 December 2010. Ensemble members: red bars; ensemble median: vertical red line; observation: vertical orange line; predictive PDFs: blue/green lines; EMOS medians: vertical blue/green lines.

based on covariate information. However, the applicability of this approach is subject to the availability of suitable covariates. For some weather variables, full mixture EMOS models can be formulated where the parameters and weights of a mixture of two forecast distributions are estimated jointly (Baran and Lerch, 2016). However, such approaches are limited to specific weather variables, and the estimation is computationally demanding.

Recently a more generally applicable route towards improving the forecast performance has received significant interest (see e.g. Möller and Groß, 2016; Yang *et al.*, 2017; Bassetti *et al.*, 2018), which is based on a two-step combination of predictive distributions from individual post-processing models. In the first step, individual EMOS models based on single parametric distributions are estimated, whereas in the second step the forecast distributions are combined utilizing state of the art forecast combination techniques such as the (spread-adjusted) linear pool, the beta-transformed linear pool (Gneiting and Ranjan, 2013), or the recently proposed Bayesian, essentially non-parametric calibration approach (Bassetti *et al.*, 2018). Besides these techniques Baran and Lerch (2018) propose a computationally efficient 'plug-in' approach to determining combination weights in the linear pool that is specific to post-processing applications.

Besides the calibration of univariate weather quantities an increasing interest has appeared in modelling correlations between the different weather variables. In the special case of wind vectors, Pinson (2012) suggested an adaptive calibration technique, whereas Schuhen *et al.* (2012) and Sloughter *et al.* (2013) introduced bivariate EMOS and BMA models, respectively. Further, Möller *et al.* (2013) developed a general approach where after univariate calibration of the weather variables, the component predictive PDFs are joined into a multivariate predictive density with the help of a Gaussian copula. Another idea appears in the ensemble copula coupling method Schefzik *et al.* (2013), where after univariate calibration the rank order information in the raw ensemble is used to restore correlations. For joint post-processing of ensemble forecasts of wind speed and temperature Baran and Möller (2015) and Baran and Möller (2017) propose bivariate

BMA and EMOS models, respectively, and finally, Schefzik (2016a,b) introduces non-parametric approaches for modelling spatial dependencies between individual univariate and multivariate post-processed forecasts.

Statistical calibration can also be applied to improve the performance of hydrological forecasts. EMOS based statistical post-processing turned out to improve the predictive performance of hydrological ensemble forecasts for different gauges along river Rhine (Hemri *et al.*, 2015; Hemri and Klein, 2017), whereas Baran *et al.* (2019a) propose a doubly truncated BMA model for calibration of Box-Cox transformed ensemble forecasts of water levels.

1.3 Post-processing approaches

As mentioned in Section 1.2, the Bayesian model averaging and ensemble model output statistics are among the most popular post-processing approaches as they provide full predictive distributions. In the present work we also concentrate on various versions of these techniques describing new models and applications.

In what follows, let $f_1, f_2, \ldots, f_K$ denote the ensemble forecast of a given weather or hydrological quantity X for a given location, time and lead time under the assumption that the ensemble members can be clearly distinguished and they are not exchangeable. Such forecasts are usually outputs of multi-model, multi-analyses EPSs, where each member can be identified and tracked. This property holds e.g. for the UWME or for the COSMO-DE ensemble.

However, recently most operational EPSs incorporate ensembles where at least some members can be considered as statistically indistinguishable and in this way exchangeable, as these forecasts are generated using perturbed initial conditions. This is the case with the 51-member operational ECMWF ensemble or one can mention multi-model EPSs such as the Grand Limited Area Model Ensemble Prediction System (GLAMEPS) ensemble (Iversen *et al.*, 2011) or the THORPEX[2] Interactive Grand Global Ensemble (Swinbank et al., 2016).

In the remaining part of this chapter, if we have M ensemble members divided into K exchangeable groups, where the kth group contains $M_k \geq 1$ ensemble members ($\sum_{k=1}^{K} M_k = M$), notation $f_{k,\ell}$ is used for the ℓth member of the kth group.

1.3.1 Bayesian model averaging

The BMA predictive distribution of a weather or hydrological quantity X for a given location, time and lead time proposed by Raftery *et al.* (2005) is a weighted mixture with PDF

$$p(x|\, f_1, \ldots, f_K; \theta_1, \ldots, \theta_K) := \sum_{k=1}^{K} \omega_k g\big(x|\, f_k, \theta_k\big), \tag{1.3.1}$$

where $g\big(x|f_k, \theta_k\big)$ is the component PDF from a parametric family corresponding to the kth ensemble member f_k with parameter (vector) θ_k to be estimated, and ω_k is the corresponding weight determined by the relative performance of this particular member

[2]The Observing System Research and Predictability Experiment

during the training period. Note that the weights should form a probability distribution, that is $\omega_k \geq 0,\ k = 1, 2, \ldots, K,$ and $\sum_{k=1}^{K} \omega_k = 1$.

To account for the existence of groups of exchangeable ensemble members, Fraley *et al.* (2010) suggest to use the same weights and parameters within a given group. Thus, if we have M ensemble members divided into K exchangeable groups, model (1.3.1) is replaced by

$$p(x|f_{1,1}, \ldots, f_{1,M_1}, \ldots, f_{K,1}, \ldots, f_{K,M_K}; \theta_1, \ldots, \theta_K) := \sum_{k=1}^{K} \sum_{\ell=1}^{M_k} \omega_k g\big(x|\, f_{k,\ell}, \theta_k\big). \qquad (1.3.2)$$

For the sake of simplicity, in Sections 2.1, 3.1, 4.1 and 5.1 we provide results and formulae only for model (1.3.1) as their extension to model (1.3.2) is rather straightforward.

Model parameters θ_k and weights $\omega_k,\ k = 1, 2, \ldots, K,$ are usually estimated using rolling training data consisting of ensemble members and verifying observations from the preceding n days. In general, a maximum likelihood approach is applied.

BMA models corresponding to various weather or hydrological quantities differ in the component parametric distribution families and in the way the parameters are linked to the ensemble members. E.g. to model temperature and sea level pressure Raftery *et al.* (2005) propose a normal mixture with predictive distribution of the form

$$\sum_{k=1}^{K} \omega_k \mathcal{N}\big(\alpha_k + \beta_k f_k, \sigma^2\big),$$

other currently available models with the corresponding quantities to be forecast are listed below.

- Wind speed:
  - Gamma mixture (Sloughter *et al.*, 2010), for details see Section 3.1.1;
  - Truncated normal mixture with cut-off at 0 from below (Baran, 2014), for details see Section 3.1.2.
- Precipitation accumulation:
  - Discrete-continuous model. Point mass at zero, gamma mixture for modelling positive precipitation accumulation (Sloughter *et al.*, 2007), for details see Section 4.1.
- Wind direction:
  - Von-Mises mixture (Bao *et al.*, 2010).
- Box-Cox transformed water levels:
  - Doubly truncated normal mixture (Baran *et al.*, 2019a), for details see Section 2.1
- Wind vector:
  - Bivariate normal mixture (Sloughter *et al.*, 2013).

- Wind speed and temperature:
  - Bivariate normal mixture truncated from below at zero in the wind coordinate (Baran and Möller, 2015), for details see Section 5.1.

1.3.2 Ensemble model output statistics

In contrast to the BMA approach, the EMOS forecast distribution is given by a single parametric law with parameters that depend on the ensemble forecast. For a given weather or hydrological quantity X for a given location, time and lead time it has the general form

$$X \mid f_1, \ldots, f_K \sim h(x| f_1, \ldots, f_K; \theta),$$

where the parametric PDF $h(x| f_1, \ldots, f_K; \theta)$ is connected to the ensemble members with the help of suitable link functions. E.g. the EMOS predictive distribution for temperature and sea level pressure suggested by Gneiting *et al.* (2005) is

$$\mathcal{N}\big(a_0 + a_1 f_1 + \ldots + a_K f_K, b_0 + b_1 S^2\big) \qquad \text{with} \qquad S^2 := \frac{1}{K-1} \sum_{k=1}^{K} \big(f_k - \overline{f}\big)^2, \quad (1.3.3)$$

where $\overline{f}$ denotes the ensemble mean.

If the ensemble contains groups of statistically indistinguishable ensemble members, members within a given group should share the same parameters (Gneiting, 2014) resulting in the exchangeable version

$$\mathcal{N}\big(a_0 + a_1 \overline{f}_1 + \cdots + a_K \overline{f}_K, b_0 + b_1 S^2\big)$$

of model (1.3.3), where $\overline{f}_k$ denotes the mean of the kth group.

Parameters of an EMOS model are estimated by optimizing the mean value of a proper scoring rule (see Section 1.4) over the forecast cases in the (usually rolling) training data.

Again, different weather or hydrological quantities require different predictive distributions and link functions.

- Wind speed:
  - Truncated normal distribution with cut-off at 0 from below (Thorarinsdottir and Gneiting, 2010), for details see Section 3.2.1;
  - Generalized extreme value distribution (Lerch and Thorarinsdottir, 2013), for details see Section 3.2.3;
  - Log-normal distribution (Baran and Lerch, 2015), for details see Section 3.2.2.
- Precipitation accumulation:
  - Censored generalized extreme value distribution (Scheuerer, 2014), for details see Section 4.2.2;
  - Censored, shifted gamma distribution (Scheuerer and Hamill, 2015; Baran and Nemoda, 2016), for details see Section 4.2.1.
- Box-Cox transformed water levels:

  - Doubly truncated normal distribution (Hemri *et al.*, 2015; Hemri and Klein, 2017), for details see Section 2.2.

- Wind vector:

  - Bivariate normal distribution (Schuhen *et al.*, 2012).

- Wind speed and temperature:

  - Bivariate normal distribution truncated from below at zero in the wind coordinate (Baran and Möller, 2017), for details see Section 5.2.

1.3.3 Parameter estimation strategies

The choice of the training data is important for statistical post-processing. As mentioned before, for estimating the BMA and EMOS model parameters usually a rolling training period is applied, and the estimates are obtained using ensemble forecasts and corresponding validating observations for the preceding n calendar days. Given a training period, there are two traditional approaches for spatial selection of the training data (Thorarinsdottir and Gneiting, 2010). In the global (regional) approach, parameters are estimated using all available forecast cases from the training period resulting in a single universal set of parameters across the entire ensemble domain. It requires quite short training periods (see e.g. Baran *et al.* (2013, 2014a,b) and Baran and Nemoda (2016), where the optimal training period lengths for ALADIN-HUNEPS wind speed, temperature and precipitation forecasts are given), but usually it is unsuitable for large and heterogeneous observation domains. For local parameter estimation, one has distinct parameter estimates for the different stations obtained using only training data of the given station. To avoid numerical stability problems, local models require much longer training periods (for optimal training period lengths for EMOS modelling of different weather quantities see e.g. Hemri *et al.*, 2014), but if the training data is large enough, it will usually outperform the regional approach. To combine the advantages of local and regional estimation, Lerch and Baran (2017) introduced two semi-local methods where the training data for a given station is augmented with data from stations with similar characteristics. The choice of similar stations is based either on suitably defined distance functions or on clustering. In the distance based approach, which generalizes the idea of Hamill *et al.* (2008), training sets of a given station are increased by including training data from the L nearest stations, and distances are measured from historical data. In the clustering based semi-local method, the observation sites are grouped into clusters using k-means clustering of feature vectors depending both on the station climatology (observations at the given station) and the forecast errors of the raw ensemble during the training period, then a regional parameter estimation is performed within each cluster. With the help of these methods one can get reliable parameter estimates even for short training periods and the obtained models may outperform the local BMA or EMOS approaches (Lerch and Baran, 2017). A detailed description of semi-local approaches to parameter estimation can be found in Chapter 6.

1.4 Forecast evaluation

In probabilistic forecasting the general aim is to access the maximal sharpness of the predictive distribution subject to calibration (Gneiting *et al.*, 2007), where the latter means a statistical consistency between the predictive distributions and the validating observations, whereas the former refers to the concentration of the predictive distribution.

One of the simplest tools for getting a first impression about the calibration of ensemble forecasts is the verification rank histogram (or Talagrand diagram), defined as the histogram of ranks of validating observations with respect to the corresponding ensemble forecasts (see e.g. Wilks, 2011, Section 8.7.2). In the case of a properly calibrated K-member ensemble, the ranks follow a uniform distribution on $\{1, 2, \ldots, K+1\}$, and the deviation from uniformity can be quantified by the reliability index Δ defined by

$$\Delta := \sum_{r=1}^{K+1} \Big| \rho_r - \frac{1}{K+1} \Big|, \tag{1.4.1}$$

where ρ_r is the relative frequency of rank r (Delle Monache *et al.*, 2006). The verification rank histogram can also be generalized to multivariate ensemble forecasts, however, in this case the usual problem is the proper definition of ranks. In Chapter 5 of the present work we use the multivariate ordering proposed by Gneiting *et al.* (2008). For a probabilistic forecast one can calculate the reliability index (and plot the verification rank histogram as well) from a preferably large number of ensembles sampled from the predictive PDF and the corresponding verifying observations.

In the univariate case the continuous counterpart of the verification rank histogram is the probability integral transform (PIT) histogram. By definition, the PIT is the value of predictive cumulative distribution function (CDF) at the validating observation (Raftery *et al.*, 2005), which in case of proper calibration should follow a uniform distribution on the $[0, 1]$ interval. Apart from the visual inspection of PIT histograms, formal statistical test of uniformity can be used to assess calibration. The simplest idea is to make use of the Kolmogorov-Smirnov test. This approach is followed in the case study of Section 2.3. However, as the PIT values of multi-step ahead probabilistic forecast exhibit serial correlation (see e.g. Diebold *et al.*, 1998) and the probabilistic forecasts cannot be assumed to be independent in space and time, one can employ a moment-based test of uniformity proposed by Knüppel (2015), which accounts for dependence in the PIT values. In particular, in Sections 3.3 and 4.3 we use the α^0_{1234} test of Knüppel (2015) that has been demonstrated to have superior size and power properties compared with alternative choices.

Predictive performance can be quantified with the help of scoring rules, which are loss functions $\mathcal{S}(F, x)$ assigning numerical values to pairs (F, x) of forecasts and observations. In the atmospheric sciences the most popular scoring rules are the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007; Wilks, 2011) and the logarithmic score (LogS; Good, 1952). For a (predictive) CDF $F(y)$ and real value (observation) x the CRPS is defined as

$$\begin{aligned} \mathrm{CRPS}\big(F, x\big) := \int_{-\infty}^{\infty} \big(F(y) - \mathbb{I}_{\{y \geq x\}}\big)^2 \mathrm{d}y &= \int_{-\infty}^{x} F^2(y) \mathrm{d}y + \int_{x}^{\infty} \big(1 - F(y)\big)^2 \mathrm{d}y \\ &= \mathsf{E}|X - x| - \frac{1}{2}\mathsf{E}|X - X'|, \end{aligned} \tag{1.4.2}$$

where $\mathbb{I}_H$ denotes the indicator of a set H, whereas X and X' are independent random variables with CDF F and finite first moment. The last representation in (1.4.2) implies that the CRPS can be expressed in the same unit as the observation. In most applications the CRPS has a simple closed form (see e.g. the `R` package `scoringRules`; Jordan *et al.*, 2017), otherwise the second integral expression in the definition (1.4.2) should be evaluated numerically. According to our tests with the mixture EMOS model of Section 3.2.5 (see also Baran and Lerch, 2016), this approach results in slightly more accurate results and faster calculations than the numerical evaluation of the first integral defining the CRPS. The logarithmic score is the negative logarithm of the predictive density $f(y)$ evaluated at the verifying observation, i.e.,

$$\mathrm{LogS}(F, x) := -\log(f(x)). \tag{1.4.3}$$

Both CRPS and LogS are proper scoring rules (Gneiting and Raftery, 2007) which are negatively oriented, that is, smaller scores indicate better forecasts.

A direct multivariate extension of the CRPS is the energy score (ES) introduced by Gneiting and Raftery (2007). Given a CDF F on $\mathbb{R}^d$ and a d-dimensional vector $\boldsymbol{x}$, the energy score is defined as

$$\mathrm{ES}(F, \boldsymbol{x}) := \mathsf{E}\|\boldsymbol{X} - \boldsymbol{x}\| - \frac{1}{2}\mathsf{E}\|\boldsymbol{X} - \boldsymbol{X}'\|, \tag{1.4.4}$$

where $\|\cdot\|$ denotes the Euclidean distance and, similar to the univariate case, $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors having distribution F. However, in most cases (see Section 5.4) the ES cannot be given in a closed form, so it is replaced by a Monte Carlo approximation

$$\widehat{\mathrm{ES}}(F, \boldsymbol{x}) := \frac{1}{n}\sum_{j=1}^{n}\|\boldsymbol{X}_j - \boldsymbol{x}\| - \frac{1}{2(n-1)}\sum_{j=1}^{n-1}\|\boldsymbol{X}_j - \boldsymbol{X}_{j+1}\|, \tag{1.4.5}$$

where $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a large random sample from F (Gneiting *et al.*, 2008). Finally, if F is a CDF corresponding to a forecast ensemble $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K$ then (1.4.5) reduces to

$$\mathrm{ES}(F, \boldsymbol{x}) = \frac{1}{K}\sum_{j=1}^{K}\|\boldsymbol{f}_j - \boldsymbol{x}\| - \frac{1}{2K^2}\sum_{j=1}^{K}\sum_{k=1}^{K}\|\boldsymbol{f}_j - \boldsymbol{f}_k\|. \tag{1.4.6}$$

Obviously, for univariate quantities (1.4.5) and (1.4.6) result in the approximation of the CRPS of a probabilistic forecast and CRPS of the raw ensemble, respectively.

Besides the CRPS one can also consider Brier scores (BS; Wilks, 2011, Section 8.4.2) for the dichotomous event that the observation x exceeds a given threshold y. For a predictive CDF $F(y)$ the Brier score is defined as

$$\mathrm{BS}\left(F, x; y\right) := \left(F(y) - \mathbb{I}_{\{y \geq x\}}\right)^2, \tag{1.4.7}$$

(see e.g. Gneiting and Ranjan, 2011), and note that the CRPS is the integral of the BS over all possible thresholds. Brier score is also negatively oriented and it plays an important role e.g. in evaluating forecasts of the probability of no precipitation.

To evaluate the goodness of fit of probabilistic forecasts to extreme values of the univariate weather quantity at hand, a useful tool to be considered is the threshold-weighted continuous ranked probability score (twCRPS)

$$\text{twCRPS}\left(F,x\right) := \int_{-\infty}^{\infty} \left(F(y) - \mathbb{I}_{\{y\geq x\}}\right)^2 \omega(y)\mathrm{d}y \tag{1.4.8}$$

introduced by Gneiting and Ranjan (2011), where $\omega(y) \geq 0$ is a weight function. Obviously, $\omega(y) \equiv 1$ corresponds to the traditional CRPS defined by (1.4.2), while to address values of the studied weather variable above a given threshold r one may set $\omega(y) = \mathbb{I}_{\{y\geq r\}}$. In the case studies of Chapter 3 we consider threshold values corresponding approximately to the 90th, 95th and 98th percentiles of the wind speed observations.

In case studies, with respect to a given score $\mathcal{S}(F,x)$, competing forecast methods can be compared by the mean score value

$$\overline{\mathcal{S}}_F := \frac{1}{N}\sum_{i=1}^{N} \mathcal{S}\big(F_i, x_i\big) \tag{1.4.9}$$

over all pairs (F_i, x_i), $i = 1, 2, \ldots, N$, of forecasts and observations in the verification data. Further, the improvement in a score $\mathcal{S}_F$ for a forecast F with respect to a reference forecast F_{ref} can be quantified with the help of the corresponding skill score (Gneiting and Raftery, 2007), defined as

$$\mathcal{S}_F^{skill} := 1 - \frac{\overline{\mathcal{S}}_F}{\overline{\mathcal{S}}_{F_{ref}}}, \tag{1.4.10}$$

where $\overline{\mathcal{S}}_{F_{ref}}$ denotes the mean score value corresponding to the reference approach. Thus, besides the CRPS, BS and twCRPS one can also investigate the continuous ranked probability skill score (CRPSS; see e.g. Murphy, 1973; Gneiting and Raftery, 2007), the Brier skill score (BSS; see e.g. Friedrichs and Thorarinsdottir, 2012) and the threshold-weighted continuous ranked probability skill score (twCRPSS; see e.g. Lerch and Thorarinsdottir, 2013), respectively. These scores are positively oriented, that is the larger the better. In the case studies of Chapters 2 and 4 we use the raw ensemble as a reference, whereas in Chapter 3 skill scores with respect to the TN EMOS model are reported.

To compare the calibration of probabilities of a dichotomous event of exceeding a given threshold calculated from the raw ensemble and the BMA and EMOS predictive distributions, one can make use of reliability diagrams (Wilks, 2011, Section 8.4.4). The reliability diagram plots the a graph of the observed frequency of the event against the binned forecast frequencies and in the ideal case this graph should lie on the main diagonal of the unit square. In the case studies of Chapter 4 the same thresholds as for the BSs are considered, whereas the unit interval is divided into 11 bins with break points $0.05, 0.15, 0.25, \ldots, 0.95$. Following Bröcker and Smith (2007) and Scheuerer (2014), the observed relative frequency of a bin is plotted against the mean of the corresponding probabilities, and inset histograms displaying the frequencies of the different bins on log 10 scales are also added.

Calibration and sharpness of a univariate predictive distribution can also be investigated using the coverage and average width of the $(1-\alpha)100\,\%$, $\alpha \in (0,1)$, central prediction interval, respectively. As coverage we consider the proportion of validating

observations located between the lower and upper $\alpha/2$ quantiles of the predictive CDF, and level α should be chosen to match the nominal coverage of the raw ensemble, that is $(K-1)/(K+1)100\%$, where again K is the ensemble size. As the coverage of a calibrated predictive distribution should be around $(1-\alpha)100\,\%$, such a choice of α allows direct comparison with the raw ensemble.

However, sharpness of an ensemble forecast or of a predictive distribution can also be quantified by its standard deviation. An obvious generalization of this idea to d-dimensional quantities is the determinant sharpness (DS; Möller *et al.*, 2013) defined as

$$\mathrm{DS} := \big(\det(\Sigma)\big)^{1/(2d)}, \tag{1.4.11}$$

where Σ is the covariance matrix of an ensemble or of a predictive PDF.

As point forecasts one can consider median and mean of the raw ensemble and of the calibrated predictive distribution, which in the univariate case are evaluated with the use of mean absolute errors (MAEs) and root mean square errors (RMSEs). Note that MAE optimal for the median, while RMSE is optimal for the mean forecasts (Gneiting, 2011; Pinson and Hagedorn, 2012). For multivariate point forecasts the RMSE should be replaced by the mean Euclidean error (EE) of forecasts from the corresponding validating observations, where the ensemble median can be obtained using the Newton-type algorithm given in Dennis and Schnabel (1983), the algorithm of Vardi and Zhang (2000), or any other method implemented, e.g. in the R package `pcaPP` (Fritz *et al.*, 2012). For a predictive distribution F one may apply the same algorithm on a preferably large sample from F.

Finally, as suggested by Gneiting and Ranjan (2011), statistical significance of the differences between the verification scores is assessed by utilizing the Diebold-Mariano (DM; Diebold and Mariano, 1995) test, which allows accounting for the temporal dependencies in the forecast errors. Given a scoring rule $\mathcal{S}$ and two competing probabilistic forecasts F and G, let

$$d_i(F,G) := \mathcal{S}(F_i, x_i) - \mathcal{S}(G_i, x_i), \qquad i = 1, 2, \ldots, N,$$

denote the score differences over the verification data of size N. The test statistic of the DM test is given by

$$t_N = \sqrt{N}\frac{\overline{\mathcal{S}}_F - \overline{\mathcal{S}}_G}{\widehat{\sigma}_N}, \tag{1.4.12}$$

where $\overline{\mathcal{S}}_F$ and $\overline{\mathcal{S}}_G$ are the mean scores (1.4.9) corresponding to forecasts F and G, respectively, and $\widehat{\sigma}_N$ is a suitable estimator of the asymptotic standard deviation of the sequence of score differences $d_i(F,G)$. Under some weak regularity assumptions, t_N asymptotically follows a standard normal distribution under the null hypothesis of equal predictive performance. Negative values of t_N indicate a better predictive performance of F, whereas G is preferred in case of positive values of t_N. In the case studies of Chapters 2–5, following suggestions of Diebold and Mariano (1995) and Gneiting and Ranjan (2011), as an estimator $\widehat{\sigma}_N$ in (1.4.12) for h step ahead forecasts we use the sample autocovariance up to lag $h-1$.

Chapter 2

Statistical post-processing of hydrological ensemble forecasts

Hydrological forecasts are important for a heterogeneous group of users such as, for instance, the operators of hydrological power plants, flood prevention authorities, or shipping companies. For rational decision making based on cost-benefit analyses an estimate of the predictive uncertainty (Krzysztofowicz, 1999; Todini, 2008) needs to be provided with any forecast. The state of the art approach of using a set of parallel runs of a hydrological model driven by meteorological ensemble forecasts provided by NWP models (Cloke and Pappenberger, 2009) gives a first estimate of the meteorological input uncertainty. However, as mentioned in Section 1.1, NWP ensembles are usually biased and underdispersed and other sources of uncertainty like hydrological model formulation, boundary and initial condition uncertainty as well as measurement uncertainties are typically neglected. Hence, statistical post-processing is important in order to reduce systematic errors and to obtain an appropriate estimate of the predictive uncertainty. In this chapter, which is based on Baran *et al.* (2019a), we introduce a novel BMA approach to post-processing hydrological ensemble forecast and in a case study dealing with water levels at gauge Kaub of river Rhine, the forecast skill of this new model is compared with the predictive performance of the recently developed EMOS method of Hemri and Klein (2017) and the raw ensemble forecasts.

2.1 Doubly truncated normal BMA model

For weather variables such as temperature or pressure, BMA models with Gaussian components provide a reasonable fit (Raftery *et al.*, 2005; Fraley *et al.*, 2010), however, water levels are typically non-Gaussian (see e.g. Duan *et al.*, 2007), moreover, they are bounded both from below and from above. These constraints should also be taken into account during model formulation. A general procedure is to normalize the forecasts and observations using, for instance, Box-Cox transformation

$$h_\lambda(x) := \begin{cases} \big(x^\lambda - 1\big)/\lambda, & \lambda \neq 0, \\ \log(x), & \lambda = 0 \end{cases} \tag{2.1.1}$$

with some coefficient λ, perform post-processing, and then back-transform the results using the inverse Box-Cox transformation (Duan *et al.*, 2007; Hemri *et al.*, 2013, 2014,

2015).

2.1.1 Model formulation

In Duan *et al.* (2007) and Hemri *et al.* (2013) the Box-Cox transformation is used prior to applying BMA in order to achieve approximate normality despite the positive skewness of water levels. Additionally, it is important to ensure that the resulting water level quantiles of the predictive distribution are within realistic physical bounds. At the upper bound of the distribution water levels should be lower than a water level threshold resulting from an extreme flood with a small exceedance probability, at the lower bound water levels should be higher than a water level threshold resulting from an extreme long-lasting low water period with a small non-exceedance probability. In order to ensure realistic values while still being able to benefit from the mathematical simplicity of Gaussian models, following the ideas of Hemri and Klein (2017), for modelling Box-Cox transformed water levels we use a doubly truncated normal distribution $\mathcal{N}_a^b\big(\mu, \sigma^2\big)$, with PDF

$$g_{a,b}\big(x|\,\mu,\sigma\big) := \frac{\frac{1}{\sigma}\varphi\big(\frac{x-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}, \quad x \in [a,b], \tag{2.1.2}$$

and $g_{a,b}\big(x|\,\mu,\sigma\big) := 0$, otherwise, where a and b are the lower and upper bounds and φ and Φ denote the PDF and the CDF of the standard normal distribution, respectively. Note that the mean and variance of $\mathcal{N}_a^b\big(\mu, \sigma^2\big)$ are

$$\begin{aligned} \kappa &= \mu + \sigma\frac{\varphi\big(\frac{a-\mu}{\sigma}\big) - \varphi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} \qquad \text{and} \\ \varrho^2 &= \sigma^2\left(1 + \frac{\frac{a-\mu}{\sigma}\varphi\big(\frac{a-\mu}{\sigma}\big) - \frac{b-\mu}{\sigma}\varphi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} - \left(\frac{\varphi\big(\frac{a-\mu}{\sigma}\big) - \varphi\big(\frac{b-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)}\right)^2\right), \end{aligned} \tag{2.1.3}$$

respectively. The proposed BMA predictive PDF (Baran *et al.*, 2019a) is

$$p\big(x|\,f_1,\ldots,f_K;\beta_{01},\ldots,\beta_{0K};\beta_{11},\ldots,\beta_{1K};\sigma\big) = \sum_{k=1}^{K}\omega_k g_{a,b}\big(x|\,\beta_{0k}+\beta_{1k}f_k,\sigma\big), \tag{2.1.4}$$

where we assume that the location of the kth mixture component is an affine function of the corresponding ensemble member f_k, and scale parameters are assumed to be equal for all component PDFs. The latter assumption is for the sake of simplicity and is common in BMA modelling (see e.g. Raftery *et al.*, 2005), whereas the form of the location parameter is in line with the truncated normal BMA model of Baran (2014), see also Section 3.1.2. Further, note that the EMOS model of Hemri and Klein (2017) (see Section 2.2) also links the ensemble members to the location and scale of the truncated normal distribution and not to the corresponding mean and variance.

2.1.2 Parameter estimation

Location parameters β_{0k}, β_{1k}, weights ω_k, $k = 1, 2, \ldots, K$, and scale parameter σ can be estimated from training data, which consists, for instance, of ensemble members and

validating observations from the preceding n days. In the BMA approach, estimates of location parameters are typically obtained by regressing the validating observations on the ensemble members, whereas weights and scale parameter(s) are obtained via ML estimation (see e.g. Raftery *et al.*, 2005; Sloughter *et al.*, 2007, 2010), where the log-likelihood function of the training data is maximized using the EM algorithm for mixture distributions (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997). However, the regression approach assumes the location parameters to be simple functions of the mean, which is obviously not the case for the truncated normal distribution, see (2.1.3). Hence, we propose a *pure ML* method estimating all model parameters by maximizing the likelihood function, which idea has already been considered e.g. by Sloughter *et al.* (2010).

In what follows, for a given location $s \in \mathcal{S}$ and time $t \in \mathcal{T}$ let $f_{k,s,t}$ denote the kth ensemble member, and denote by $x_{s,t}$ the corresponding validating observation. Here $\mathcal{S}$ denotes the set of locations sharing the same BMA model parameters and $\mathcal{T}$ is the set of training dates. In the case study of Section 2.3, $\mathcal{S}$ consists of a single location, however, for more complex ensemble domains different choices of training data are possible, for more details see Chapter 6. Further, as in the case study of Section 2.3 the different lead times are treated separately, reference to the lead time of the forecast is omitted. By assuming the conditional independence of forecast errors with respect to the ensemble members in space and time, the log-likelihood function for model (2.1.4) corresponding to all forecast cases (s,t) in the training set equals

$$\ell(\omega_1,\ldots,\omega_K,\beta_{01},\ldots,\beta_{0K},\beta_{11},\ldots,\beta_{1K},\sigma) = \sum_{s,t} \log\left[\sum_{k=1}^{K} \omega_k g_{a,b}\big(x_{s,t}|\,\beta_{0k}+\beta_{1k}f_{k,s,t},\sigma\big)\right]. \tag{2.1.5}$$

To obtain the ML estimates we apply EM algorithm for truncated Gaussian mixtures proposed by Lee and Scott (2012) with a mean correction. In line with the classical EM algorithm for mixtures (McLachlan and Krishnan, 1997), first we introduce latent binary indicator variables $z_{k,s,t}$ identifying the mixture component where the observation $x_{s,t}$ comes from, that is $z_{k,s,t}$ is one or zero according as whether $x_{s,t}$ follows or not the kth component distribution. Using these indicator variables one can provide the complete data log-likelihood corresponding to (2.1.5) in the form

$$\begin{aligned}\ell_C(\omega_1,\ldots,\omega_K,\beta_{01},\ldots,\beta_{0K},\beta_{11},\ldots,\beta_{1K},\sigma) \qquad (2.1.6)\\ = \sum_{s,t}\sum_{k=1}^{K} z_{k,s,t}\left[\log\big(\omega_k\big) + \log\Big(g_{a,b}\big(x_{s,t}|\,\mu_{k,s,t},\sigma\big)\Big)\right],\end{aligned}$$

with $\mu_{k,s,t} := \beta_{0k} + \beta_{1k}f_{k,s,t}$. After specifying the initial values of the parameters the EM algorithm alternates between an expectation (E) and a maximization (M) step until convergence. As first guesses $\beta_{0k}^{(0)}$ and $\beta_{1k}^{(0)}$, $k = 1,2,\ldots,K$, for the location parameters we suggest to use the coefficients of the linear regression of $x_{s,t}$ on $f_{k,s,t}$, so $\mu_{k,s,t}^{(0)} = \beta_{0k}^{(0)} + \beta_{1k}^{(0)} f_{k,s,t}$. Initial scale $\sigma^{(0)}$ can be the standard deviation of the observations in the training data set or the average residual standard deviation from the above regression, whereas the initial weights might be chosen uniformly, that is $\omega_k^{(0)} = 1/K$, $k = 1,2,\ldots,K$. Then in the E step the latent variables are estimated using the conditional expectation of the complete log-likelihood on the observed data, while in the M step the parameter estimates are updated by maximizing ℓ_C given the actual values of the latent variables.

For the doubly truncated normal model specified by (2.1.2) and (2.1.4), the E step of the $(j+1)$st iteration is

$$z_{k,s,t}^{(j+1)} := \frac{\omega_k^{(j)} g_{a,b}\big(x_{s,t}|\,\mu_{k,s,t}^{(j)}, \sigma^{(j)}\big)}{\sum_{i=1}^{K} \omega_i^{(j)} g_{a,b}\big(x_{s,t}|\,\mu_{i,s,t}^{(j)}, \sigma^{(j)}\big)}. \tag{2.1.7}$$

Once the estimates of the indicator variables (which are not necessary 0 or 1 any more) are given, the first part of the M step updating the weights is obviously

$$\omega_k^{(j+1)} := \frac{1}{N} \sum_{s,t} z_{k,s,t}^{(j+1)}, \tag{2.1.8}$$

where N is the total number of forecast cases in the training set.

Further, non-linear equations $\frac{\partial \ell_C}{\partial \beta_{0k}} = 0$ and $\frac{\partial \ell_C}{\partial \beta_{1k}} = 0$, $k = 1, 2, \ldots, K$, lead us to update formulae

$$\beta_{0k}^{(j+1)} := \left[\sum_{s,t} z_{k,s,t}^{(j+1)}\right]^{-1} \sum_{s,t} z_{k,s,t}^{(j+1)} \left\{ \big(x_{k,s,t} - \beta_{1k}^{(j)} f_{k,s,t}\big) + \sigma^{(j)} \frac{\varphi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \varphi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)}{\Phi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \Phi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)} \right\}, \tag{2.1.9}$$

$$\beta_{1k}^{(j+1)} := \left[\sum_{s,t} z_{k,s,t}^{(j+1)} f_{k,s,t}^2\right]^{-1} \sum_{s,t} z_{k,s,t}^{(j+1)} f_{k,s,t} \left\{ \big(x_{k,s,t} - \beta_{0k}^{(j)}\big) + \sigma^{(j)} \frac{\varphi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \varphi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)}{\Phi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \Phi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)} \right\},$$

respectively. However, using then simply $\mu_{k,s,t}^{(j+1)} := \beta_{0k}^{(j+1)} + \beta_{1k}^{(j+1)} f_{k,s,t}$ as the update of the location parameter results in an unstable parameter estimation process due to numerical issues. Hence, we introduce a mean correction of form

$$\mu_{k,s,t}^{(j+1)} := \mu_{k,s,t}^{(0)} - \sigma^{(j)} \frac{\varphi\Big(\frac{a-\beta_{0k}^{(j+1)} - \beta_{1k}^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\Big) - \varphi\Big(\frac{b-\beta_{0k}^{(j+1)} - \beta_{1k}^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\Big)}{\Phi\Big(\frac{b-\beta_{0k}^{(j+1)} - \beta_{1k}^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\Big) - \Phi\Big(\frac{a-\beta_{0k}^{(j+1)} - \beta_{1k}^{(j+1)} f_{k,s,t}}{\sigma^{(j)}}\Big)}, \tag{2.1.10}$$

which reflects to the difference between the location and mean of a truncated normal distribution, see (2.1.3). Finally, from $\frac{\partial \ell_C}{\partial \sigma} = 0$ we obtain the last update formula

$$\sigma^{2(j+1)} := \frac{1}{N} \sum_{s,t} \sum_{k=1}^{K} z_{k,s,t}^{(j+1)} \left\{ \big(x_{s,t} - \mu_{k,s,t}^{(j+1)}\big)^2 \right. \tag{2.1.11}$$
$$\left. + \sigma^{(j)} \frac{\big(b - \mu_{k,s,t}^{(j+1)}\big)\varphi\Big(\frac{b-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\Big) - \big(a - \mu_{k,s,t}^{(j+1)}\big)\varphi\Big(\frac{a-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\Big)}{\Phi\Big(\frac{b-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\Big) - \Phi\Big(\frac{a-\mu_{k,s,t}^{(j+1)}}{\sigma^{(j)}}\Big)} \right\}.$$

Note that without truncation $(-a = b = \infty)$ the terms of (2.1.9) and (2.1.11) depending on $\sigma^{(j)}$ disappear, so location (mean) and scale (standard deviation) are updated separately, no mean correction is required, and we get back the classical EM algorithm for normal mixtures.

As a more simple alternative approach, referred as *mean corrected,* one can omit the update step (2.1.9) for β_{0k} and β_{1k}, simplify the mean correction step (2.1.10) to

$$\mu_{k,s,t}^{(j+1)} := \mu_{k,s,t}^{(0)} - \sigma^{(j)} \frac{\varphi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \varphi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)}{\Phi\Big(\frac{b-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big) - \Phi\Big(\frac{a-\mu_{k,s,t}^{(j)}}{\sigma^{(j)}}\Big)}, \tag{2.1.12}$$

and only after the EM algorithm stops, estimate location parameters β_{0k} and β_{1k} from a linear regression of the final value of $\mu_{k,s,t}$ on $f_{k,s,t}$.

Finally, one can also try the classical *naive* approach, where location parameters β_{0k} and β_{1k} are not updated at all, that is $\mu_{k,s,t}^{(j+1)} \equiv \beta_{0k}^{(0)} + \beta_{1k}^{(0)} f_{k,s,t}$.

In the case study of Section 2.3 the latter two approaches do not show significantly different forecast skills, so only the results for the naive and pure ML BMA approaches are reported. The two simple approaches provide very similar location and scale parameters, the corresponding predictive distributions mainly differ in weights, whereas the pure ML method results in completely different location parameters.

2.2 Truncated normal EMOS model

In the EMOS approach to calibration of Box-Cox transformed ensemble forecasts of water levels proposed by Hemri and Klein (2017), the predictive distribution is a single doubly truncated normal distribution $\mathcal{N}_a^b\big(\mu, \sigma^2\big)$ defined by (2.1.2), and the ensemble members are just linked to the location μ and scale σ via equations

$$\mu = a_0 + a_1 f_1 + \cdots + a_K f_K \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 S^2, \tag{2.2.13}$$

where S^2 denotes the variance of the transformed ensemble. In the case of existence of groups of exchangeable ensemble members the equation for the location in (2.2.13) is replaced by

$$\mu = a_0 + a_1 \overline{f}_1 + \cdots + a_K \overline{f}_K, \tag{2.2.14}$$

where $\overline{f}_k$ denotes the mean value of the kth group. According to the optimum score estimation principle of Gneiting and Raftery (2007), location parameters $a_0, a_1, \ldots, a_K \in \mathbb{R}$ and scale parameters $b_0, b_1 \geq 0$ are estimated from the training data by optimizing the mean value of a proper verification score, which is usually the CRPS defined by (1.4.2). Note that for the doubly truncated normal distribution $\mathcal{N}_a^b\big(\mu, \sigma^2\big)$ the CRPS has a closed form (Jordan *et al.*, 2017), namely

$$\begin{aligned} \mathrm{CRPS}\Big(\mathcal{N}_a^b\big(\mu, \sigma^2\big), x\Big) = \sigma \Bigg[& \frac{\frac{z-\mu}{\sigma}\Big[2\Phi\big(\frac{x-\mu}{\sigma}\big) - \Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)\Big]}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} \\ & + \frac{2\varphi\big(\frac{x-\mu}{\sigma}\big)}{\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)} - \frac{\Phi\big(\frac{\sqrt{2}(b-\mu)}{\sigma}\big) - \Phi\big(\frac{\sqrt{2}(a-\mu)}{\sigma}\big)}{\sqrt{\pi}\Big[\Phi\big(\frac{b-\mu}{\sigma}\big) - \Phi\big(\frac{a-\mu}{\sigma}\big)\Big]^2} \Bigg]. \end{aligned} \tag{2.2.15}$$

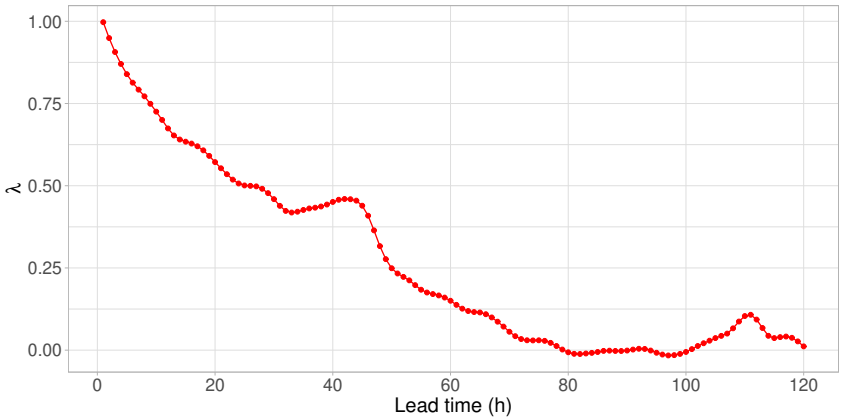


Figure 2.1: Box-Cox transformation parameter λ as function of the lead time.

2.3 Case study

The predictive performance of the truncated normal BMA approach described in Section 2.1 and its suitability for hydrological ensemble forecasts is assessed at the example of multi-model ensemble forecasts of water level at gauge Kaub at river Rhine.

2.3.1 Data

BMA and EMOS calibration approaches are tested on ensemble forecasts of water level (cm) at gauge Kaub of river Rhine (546 km) and the corresponding validating observations. Predictions for an eight year period between 1 January 2008 and 31 December 2015 are investigated with lead times from 1 h to 120 h with a time step of 1 h. The minimum and maximum recorded water levels at this particular gauge are 35 cm and 825 cm, respectively. Our 79-member multimodel water level ensemble is obtained by plugging ensemble forecasts for the relevant weather variables produced by different ensemble prediction systems into the hydrological model HBV-96 (Lindström *et al.*, 1997), which is run at the German Federal Institute of Hydrology (BfG) for operational runoff forecasting. We consider the ECMWF high resolution (HRES) forecast, the 51-member ECMWF forecast (ENS) (Molteni *et al.*, 1996; Leutbecher and Palmer, 2008), the 16-member COSMO LEPS forecast of the limited-area ensemble prediction system of the consortium for small-scale modelling (Montani *et al.*, 2011) and the 11-member NCEP GEFS forecast of the reforecast version 2 of the global ensemble forecast system of the National Center for Environmental Prediction (Hamill *et al.*, 2013). The runoff forecasts are then converted into water level forecasts for the navigation-relevant gauges, including gauge Kaub, using a hydrodynamic model. All ensemble forecast are initialized at 6 UTC. We remark that the data set at hand is part of the data studied in Hemri and Klein (2017).

2.3.2 Verification results

As mentioned in Sections 2.1 and 2.2, BMA and EMOS post-processing is applied for modelling Box-Cox transformed water levels. As in Hemri and Klein (2017), each lead

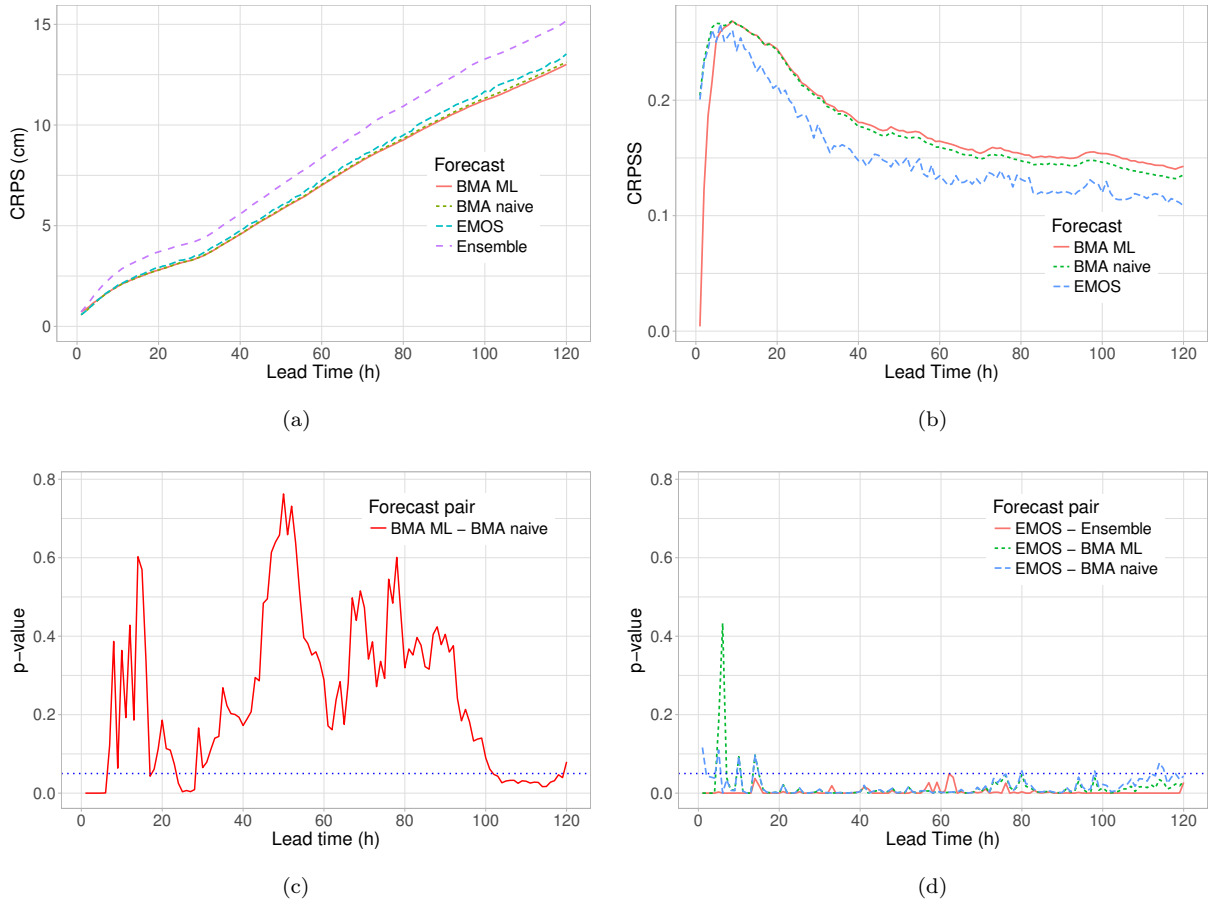
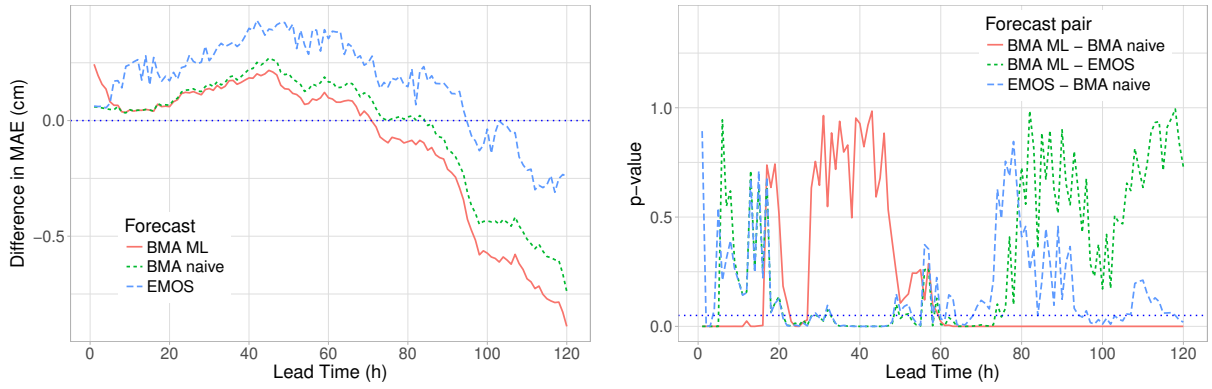


Figure 2.2: Mean CRPS values (a) and CRPSS with respect to the raw ensemble (b); p-values of DM tests for equality of mean CRPS of the two BMA approaches (c) and of all models compared with EMOS (d). Horizontal dotted lines of (c) and (d) indicate a 5 % level of significance.

time has an individual Box-Cox parameter λ (see Figure 2.1) maximizing the in-sample skill of seasonally fitted EMOS models in terms of the CRPS relative to the raw ensemble, where data from the same season of other years are used for training. These estimates are then averaged over the training periods in order to obtain one estimate per lead time. Obviously, for a given lead time the same coefficient is applied both for the forecasts and observations.

Similar to Hemri and Klein (2017), we assume that water levels are in the interval spanned by half of the minimum and double of the maximum recorded water level, i.e. they are between 17.5 cm and 1650 cm, so the Box-Cox transforms of these values serve as lower and upper bounds for the truncated normal distribution used both in BMA and EMOS modelling.

The generation of the hydrological ensemble forecast described in Section 2.3.1 induces a natural grouping of the ensemble members. One contains just the forecast based on the ECMWF HRES, the other 51-member group corresponds to the ECMWF ENS, whereas forecasts based on COSMO LEPS and NCEP GEFS ensemble weather forecasts form two other groups of sizes 16 and 11, respectively. Hence, Box-Cox transformed water level forecasts are calibrated using the truncated normal BMA model for exchangeable



(a) (b)

Figure 2.3: Difference in MAE values from the raw ensemble (a) and p-values of DM tests for equality of MAE of the various post-processing approaches (b). Horizontal dotted lines indicate the reference raw ensemble (a) and a 5 % level of significance (b).

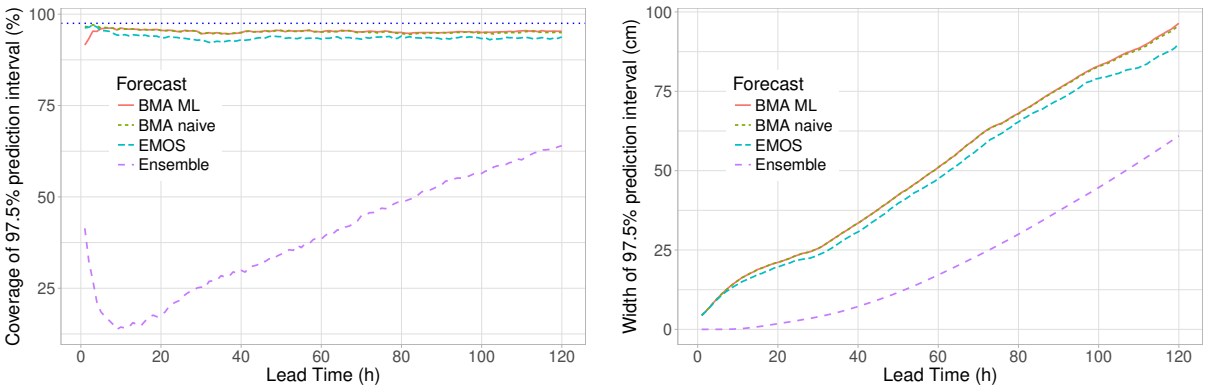
ensemble members specified by (1.3.2) and (2.1.2), and truncated normal EMOS given by (2.2.13) and (2.2.14) with $K = 4$ and $M_1 = 1$, $M_2 = 51$, $M_3 = 16$, $M_4 = 11$. This means that for BMA modelling 12, whereas for finding the EMOS predictive distribution 7 free parameters have to be estimated. To ensure a reasonably stable parameter estimation we use a rolling training period of length 100 days. Thus, BMA and EMOS models are verified on the period 10 April 2008 – 31 December 2015 (2822 calendar days). Further, we consider one day ahead calibration for all lead times. This means that for modelling water level e.g. for 1 January 2015 we use forecasts and observations for the preceding 100 days ending at 31 December 2014. For 24 h lead time the last forecasts are initialized at 30 December 2014, whereas for 120 h lead time at 26 December 2014.

While BMA and EMOS models are fit to Box-Cox transformed values $X \in [a, b]$, to ensure comparability we provide verification scores for the original forecasts and observations. This means that for quantile based scores (MAE, coverage, average width), before evaluating the score, the inverse Box-Cox transformation h_λ^{-1} is applied to the appropriate quantiles of the predictive CDF F, whereas the CRPS corresponding to the predictive CDF $G(y) := F\big(h_\lambda(y)\big)$ of the original water level $Y = h_\lambda^{-1}(X) \in \big[h_\lambda^{-1}(a), h_\lambda^{-1}(b)\big]$ and a real value y equals

$$\mathrm{CRPS}\left(G, y\right) = \int_{h_\lambda^{-1}(a)}^{y} F^2\big(h_\lambda(u)\big)\mathrm{d}u + \int_{y}^{h_\lambda^{-1}(b)} \Big(1 - F\big(h_\lambda(u)\big)\Big)^2 \mathrm{d}u,$$

which integral should be approximated numerically.

In Figure 2.2a the mean CRPS values of the different post-processing approaches and the raw ensemble are plotted as functions of the lead time. Note that compared with the raw ensemble all calibration approaches reduce the mean CRPS and the gap increases together with the lead time. The differences between the forecast skills are more pronounced in Figure 2.2b showing the CRPSS values with respect to the raw ensemble forecast. Note that all three presented methods have their maximal skill score at hour 9. This reflects that the relative gap in CRPS between raw and post-processed forecasts is increasing up to hour 9 and decreasing again thereafter. However, it does not imply that the absolute forecast skill increases with lead time between hour 1 and hour 9. For shorter



(a) (b)

Figure 2.4: Coverage (a) and average width (b) of nominal 97.5 % central prediction intervals. In panel (a) the ideal coverage is indicated by the horizontal dotted line.

lead times this increase is very fast and naive BMA shows the best predictive performance, whereas for longer lead times the pure ML BMA starts dominating. Obviously, longer lead times are also associated with larger forecast uncertainty which should be taken into account when one compares predictive performance. According to the results of DM tests for equal predictive performance, naive BMA significantly outperforms the raw ensemble for all lead times and the same holds for the pure ML BMA except hour 1. In general, in terms of the mean CRPS the two BMA approaches differ significantly mainly for very short and long lead times, as can be observed on the graph of p-values displayed in Figure 2.2c. EMOS also significantly outperforms the raw ensemble for all lead times, and except for the first couple of hours underperforms the BMA approaches, as depicted in Figure 2.2d.

There is much less variety in the performance of BMA and EMOS calibrated medians in terms of the MAE. According to Figure 2.3a showing the difference in MAE with respect to the raw ensemble the pure ML BMA has the best forecast skill, however, even this approach underperforms the raw ensemble until hour 70. Note that DM tests for equality of MAE values indicate that all differences plotted in Figure 2.3a are significant (DM test results are not reported), which will definitely not be the case if we compare the performance of the three post-processing methods, see the p-values of Figure 2.3b.

The positive effect of post-processing on calibration can be clearly observed on Figure 2.4a showing the coverages of nominal 97.5 % central prediction intervals as functions of the lead time. All post-processing approaches for all lead times result in almost perfect coverage, whereas the coverage of the raw ensemble is much lower and strongly depends on the lead time. The coverage values of the two BMA approaches are almost identical and after hour 4 they are closer to the nominal value than those of the EMOS. Finally, as depicted in Figure 2.4b, the raw ensemble produces the sharpest forecasts for all lead times, however, at the cost of being uncalibrated. This is fully in line with the verification rank histograms of the raw ensemble and PIT histograms of post-processed forecasts for lead times 24, 72 and 120 hours plotted in Figure 2.5. All verification rank histograms are strongly U-shaped (and the same holds for other lead times, not reported), indicating that the raw ensemble is strongly underdispersive and requires post-processing. BMA and EMOS approaches significantly improve the statistical calibration of the forecast and

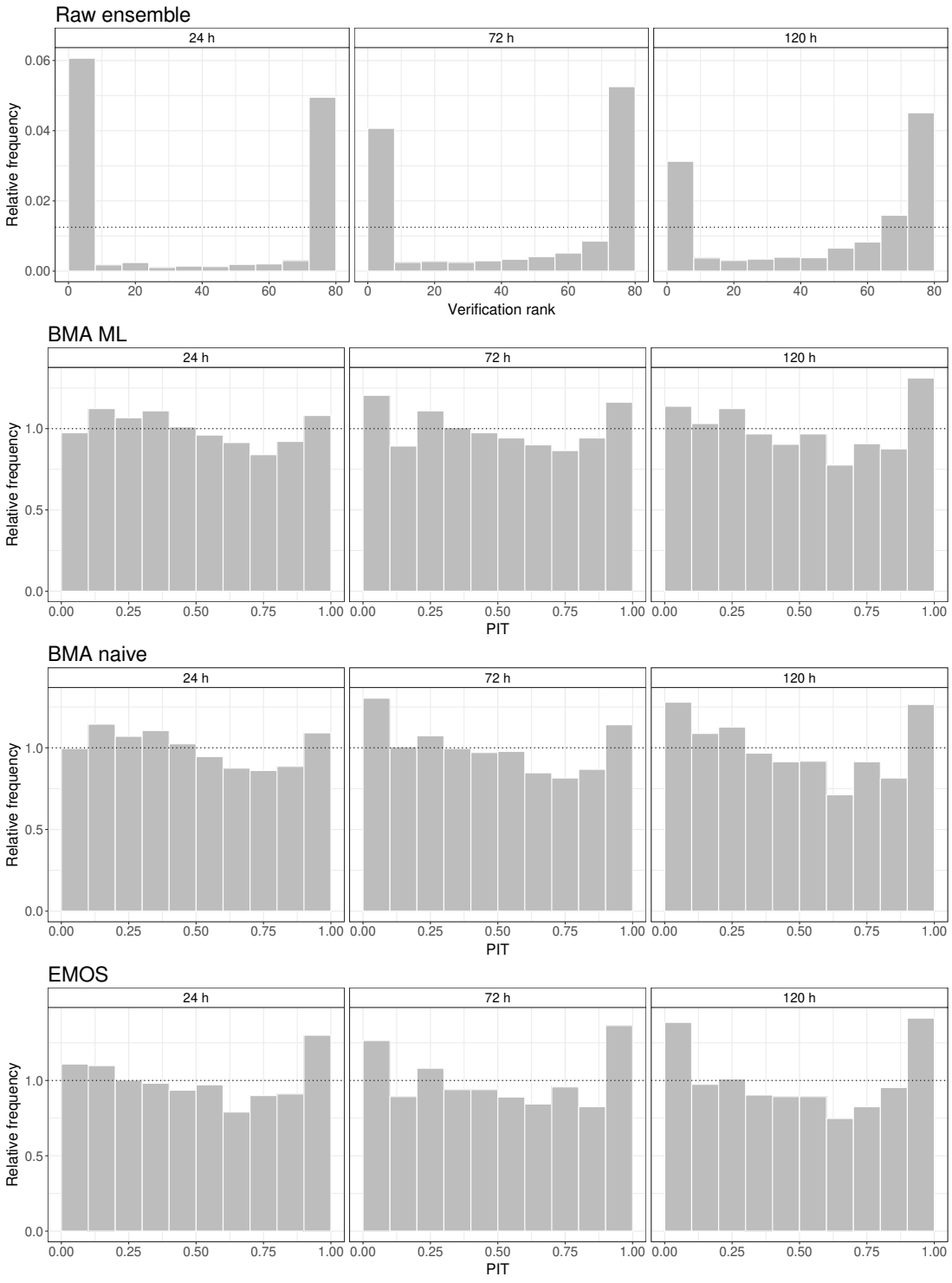


Figure 2.5: Verification rank histogram of the raw ensemble and PIT histograms of the BMA and EMOS post-processed forecasts for lead times 24, 72 and 120 hours.

result in more uniform PIT histograms, although for hour 120 naive BMA and EMOS still show a slight underdispersion. Figure 2.6 displays the values of the test statistic of the Kolmogorov-Smirnov test for uniformity of PIT values for different post-processing approaches. Although the uniformity of the PIT values of pure ML BMA, naive BMA

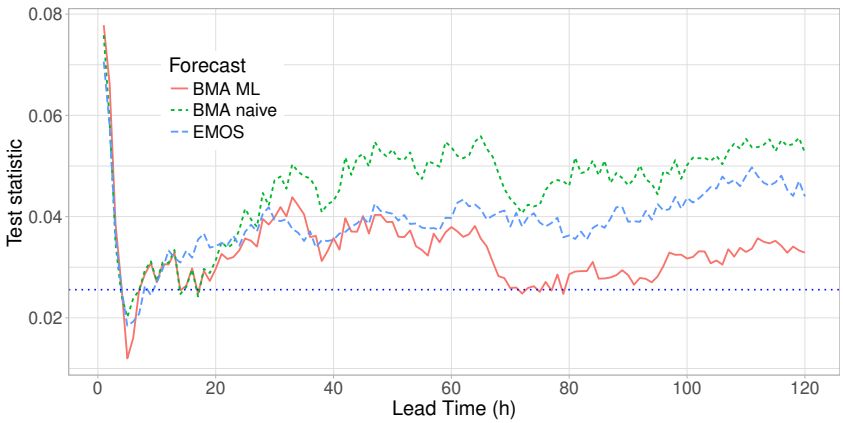


Figure 2.6: Values of the test statistic of Kolmogorov-Smirnov tests for uniformity of PIT values. Smaller values indicate better fit, dotted horizontal line corresponds to 5 % level of significance.

and EMOS can be accepted at a 5 % level of significance for only 9 (5, 6, 7, 14, 17, 72, 75, 77, 79 h), 6 (4, 5, 6, 7, 14, 17 h) and 4 (5, 6, 7, 9 h) different lead times, Figure 2.6 nicely illustrates the ranking of different approaches in terms of goodness of fit of PIT.

2.4 Conclusions

In this chapter we describe a BMA model of for calibrating Box-Cox transformed hydrological ensemble forecasts of water level, providing a predictive distribution which is a weighted mixture of doubly truncated normal distributions. The model with three different parameter estimation approaches is tested on the 79-member BfG ensemble forecast of water level at gauge Kaub of river Rhine for 120 different lead times. For verification we use the CRPS of the probabilistic forecast distributions and the MAE of the corresponding median forecasts. Further, we analyse the coverage and the average width of nominal central prediction intervals, which serve as measures of calibration and sharpness, respectively. Furthermore, the forecast skill of the BMA model is compared with that of the recently introduced EMOS model of Hemri and Klein (2017) and the raw ensemble.

Based on the results of the presented case study one can conclude that compared with the raw ensemble, post-processing always improves the calibration of probabilistic and accuracy of point forecasts. Further, BMA model utilizing pure ML for parameter estimation has the best predictive performance and, except very short lead times, the BMA approach significantly outperforms the EMOS calibration.

Chapter 3

Statistical calibration of wind speed forecasts

In our industrialized world several important applications require reliable and accurate wind speed forecasts. These include, but are not limited to agriculture, aviation or wind energy production. In particular, high wind speeds can cause severe damages to infrastructure and their predictions are important parts of weather warnings.

Nowadays, weather services typically produce ensemble forecasts for wind speed, however, these forecasts often suffer from the lack of calibration calling again for some form of post-processing (Buizza, 2018).

In this chapter we introduce several approaches to post-processing wind speed ensemble forecasts and test the predictive skill of these new methods using three different data sets containing forecasts produced by completely different ensemble prediction systems (UWME, ECMWF EPS, ALADIN-HUNEPS EPS) which cover different forecast domains. The chapter is based on Baran (2014), Baran and Lerch (2015) and Baran and Lerch (2016) and uses also some results of Baran *et al.* (2013) and Baran *et al.* (2014b).

3.1 BMA models for wind speed

To model wind speed one requires non-negative and skewed distributions. A popular candidate is the Weibull distribution (see e.g. Justus *et al.*, 1978), however, gamma distribution is also a traditional choice (Garcia *et al.*, 1988).

3.1.1 Gamma BMA model

In the BMA approach of Sloughter *et al.* (2010) for modelling wind speed the component PDFs in the predictive distribution (1.3.1) follow a gamma law $\Gamma(\kappa, \theta)$ with shape $\kappa > 0$ and scale $\theta > 0$ having PDF

$$g(x|\,\kappa, \theta) := \begin{cases} \frac{x^{\kappa-1}\mathrm{e}^{-x/\theta}}{\theta^{\kappa}\Gamma(\kappa)}, & x > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{3.1.1}$$

where $\Gamma(\kappa)$ denotes value of the gamma function at κ. A gamma distribution can also be parametrized by its mean $\mu > 0$ and standard deviation $\sigma > 0$ using expressions

$$\kappa = \mu^2/\sigma^2 \qquad \text{and} \qquad \theta = \sigma^2/\mu,$$

and Sloughter *et al.* (2010) suggests to express these quantities as affine functions

$$\mu_k = b_{0k} + b_{1k} f_k \qquad \text{and} \qquad \sigma_k = c_0 + c_1 f_k \tag{3.1.2}$$

of the corresponding ensemble member f_k. This results in a BMA predictive distribution

$$p(x|\, f_1, \ldots, f_K; b_{01}, \ldots, b_{0K}; b_{11}, \ldots, b_{1K}; c_0, c_1) = \sum_{k=1}^{K} \omega_k g_k(x|\, f_k),$$

where $g_k(x|\, f_k)$ denotes the PDF of a gamma distribution with mean and standard deviation specified by (3.1.2).

Mean parameters b_{0k}, b_{1k} are estimated from the training data using a simple linear regression of the validating observations on the corresponding ensemble members, whereas for weights ω_k and standard deviation parameters c_0 and c_1, as usual, the likelihood function is maximized with the help of the EM algorithm. However, in this situation, in contrast e.g. to the normal BMA model of Raftery *et al.* (2005), one cannot use closed formulae in the maximization step (see Section 2.1.2), so a numerical optimization is required, which increases the computation costs of modelling.

3.1.2 Truncated normal BMA model

Using the ideas of Thorarinsdottir and Gneiting (2010), for modelling wind speed we consider a mixture of truncated normal distributions with cut-off at zero (Baran, 2014). The proposed predictive distribution is

$$p\big(x|\, f_1, \ldots, f_K; \beta_{01}, \ldots, \beta_{0K}; \beta_{11}, \ldots, \beta_{1K}; \sigma\big) = \sum_{k=1}^{K} \omega_k g_{0,\infty}\big(x|\, \beta_{0k} + \beta_{1k} f_k, \sigma\big), \tag{3.1.3}$$

where $g_{a,b}\big(x|\, \mu, \sigma\big)$ is the PDF of the doubly truncated normal distribution defined by (2.1.2). In this way model (3.1.3) is a special case of the doubly truncated BMA model (2.1.4) discussed in Section 2.1 with $a = 0$ and $b = \infty$, however, one should note that the latter hydrological model was introduced four years later than the former one for wind speed.

Similar to Section 2.1.2, three different approaches to parameter estimation are considered (for detailed description see Baran, 2014).

- *Naive*: location parameters β_{0k}, β_{1k} are estimated from the training data by regressing the validating observations on the corresponding ensemble members. ML method with EM algorithm for truncated normal mixtures is used for obtaining weights ω_k and scale σ.

- *Mean corrected*: location parameters are estimated as in the naive approach, however, in each step of the EM algorithm a mean correction similar to (2.1.12) is applied to account for the difference between the mean and location parameter of a truncated normal distribution (see (2.1.3)).

- *Pure ML*: all parameters are estimated by ML, the update formulae for the EM algorithm are special cases of (2.1.7) – (2.1.11) for $a = 0$, $b = \infty$.

Note that in the EM algorithm all three methods use closed formulae.

3.2 EMOS models for wind speed

Besides the BMA models of Section 3.1, over recent years a wide range of EMOS based post-processing approaches and modelling strategies for wind speed forecasts have been introduced.

3.2.1 Truncated normal EMOS model

Consider again the TN distribution $\mathcal{N}_0^\infty\big(\mu, \sigma^2\big)$ with location μ, scale $\sigma > 0$ and cut-off at zero. The EMOS predictive distribution of wind speed proposed by Thorarinsdottir and Gneiting (2010) is

$$\mathcal{N}_0^\infty\big(a_0 + a_1 f_1 + \cdots + a_K f_K, b_0 + b_1 S^2\big) \quad \text{with} \quad S^2 := \frac{1}{K-1}\sum_{k=1}^{K}\big(f_k - \overline{f}\big)^2, \tag{3.2.4}$$

which is a special case of the hydrological EMOS model (2.2.13) discussed in Section 2.2. Further, if the ensemble can be divided into groups of exchangeable members, the predictive distribution changes to

$$\mathcal{N}_0^\infty\big(a_0 + a_1 \overline{f}_1 + \cdots + a_K \overline{f}_K, b_0 + b_1 S^2\big), \tag{3.2.5}$$

see also (2.2.14). To estimate location parameters $a_0 \in \mathbb{R}$, $a_1, \ldots, a_K \geq 0$ and scale parameters $b_0, b_1 \geq 0$, one can again use the optimum score estimation principle and minimize an appropriate verification score over the training data.

3.2.2 Log-normal EMOS model

As an alternative to the TN model of Section 3.2.1, we propose an EMOS approach based on an LN distribution (Baran and Lerch, 2015). This distribution has a heavier upper tail, and in this way it is more appropriate to model high wind speed values. The PDF of the LN distribution $\mathcal{LN}\big(\mu, \sigma\big)$ with location μ and shape $\sigma > 0$ is

$$h(x|\,\mu, \sigma) := \frac{1}{x\sigma}\varphi\big((\log x - \mu)/\sigma\big), \quad x \geq 0, \tag{3.2.6}$$

and $h(x|\,\mu, \sigma) := 0$, otherwise, while the mean m and variance v of this distribution are

$$m = \mathrm{e}^{\mu + \sigma^2/2} \qquad \text{and} \qquad v = \mathrm{e}^{2\mu+\sigma^2}\big(\mathrm{e}^{\sigma^2} - 1\big),$$

respectively. Further, since

$$\mu = \log\left(\frac{m^2}{\sqrt{v + m^2}}\right) \qquad \text{and} \qquad \sigma = \sqrt{\log\Big(1 + \frac{v}{m^2}\Big)}, \tag{3.2.7}$$

an LN distribution can also be parametrized by these quantities. In our EMOS approach m and v are affine functions of the ensemble members and ensemble variance, respectively, that is

$$m = \alpha_0 + \alpha_1 f_1 + \cdots + \alpha_K f_K \quad \text{and} \quad v = \beta_0 + \beta_1 S^2. \tag{3.2.8}$$

Similar to the TN model, to obtain the values of mean and variance parameters $\alpha_0 \in \mathbb{R}$, $\alpha_1, \ldots, \alpha_K \geq 0$ and $\beta_0, \beta_1 \geq 0$, respectively, one has to perform an optimum score

estimation based on some verification measure. Obviously, for the case of exchangeable ensemble members instead of (3.2.8) we have

$$m = \alpha_0 + \alpha_1 \overline{f}_1 + \cdots + \alpha_K \overline{f}_K \quad \text{and} \quad v = \beta_0 + \beta_1 S^2. \tag{3.2.9}$$

3.2.3 Generalized extreme value EMOS model

Another approach to post-processing of wind speed ensemble forecasts is to consider a GEV distribution $\mathcal{GEV}\big(\mu, \sigma, \xi\big)$ with location μ, scale $\sigma > 0$ and shape ξ characterized by CDF

$$H(x|\,\mu, \sigma, \xi) := \begin{cases} \exp\Big(- \big[1 + \xi(\frac{x-\mu}{\sigma})\big]^{-1/\xi}\Big), & \xi \neq 0; \\ \exp\Big(- \exp\big(- \frac{x-\mu}{\sigma}\big)\Big), & \xi = 0, \end{cases} \tag{3.2.10}$$

if $1+\xi(x-\mu)/\sigma > 0$ and zero otherwise. This definition shows the main disadvantage of using a GEV distribution for modelling wind speed: namely, there is a positive probability for a GEV distributed random variable to be negative.

For calibrating ECMWF ensemble forecasts of wind speed over Germany, Lerch and Thorarinsdottir (2013) suggest to model location and scale parameters by

$$\mu = \gamma_0 + \gamma_1 f_1 + \cdots + \gamma_K f_K \qquad \text{and} \qquad \sigma = \sigma_0 + \sigma_1 \overline{f}, \tag{3.2.11}$$

while the shape parameter ξ is considered to be independent of the ensemble. The exchangeable version of the GEV EMOS models operates with link functions

$$\mu = \gamma_0 + \gamma_1 \overline{f}_1 + \cdots + \gamma_K \overline{f}_K \qquad \text{and} \qquad \sigma = \sigma_0 + \sigma_1 \overline{f}. \tag{3.2.12}$$

In general, one can also incorporate the ensemble variance into the models of location and scale. However, preliminary studies showed that models (3.2.11) and (3.2.12) are also reasonable choices for the case studies of Section 3.3. Again, all model parameters are estimated by optimizing an appropriate verification measure over the training data.

3.2.4 Regime-switching models

To combine the advantageous properties of light-tailed (TN) and heavy-tailed (LN or GEV) approaches, one can also investigate a regime-switching method (Lerch and Thorarinsdottir, 2013). Depending on the value of the ensemble median f_{med}, one can consider either a light- or a heavy-tailed distribution based EMOS model. Given a threshold $\theta > 0$, the EMOS predictive distribution is e.g. $\mathcal{N}_0^\infty\big(\mu_{TN}, \sigma_{TN}^2\big)$ if $f_{med} < \theta$ and $\mathcal{LN}\big(\mu_{LN}, \sigma_{LN}\big)$, otherwise (Baran and Lerch, 2015). Model parameters μ_{TN} and σ_{TN} depend on the ensemble forecast according to (3.2.4) or (3.2.5), while the expressions for μ_{LN} and σ_{LN} can be obtained form (3.2.8) or (3.2.9) via transformation (3.2.7). For training the combined model, we propose two different methods. If the training data set is large enough, i.e. many forecast cases belong to each day to be investigated, the heavy-tailed (LN or GEV) model is trained using only ensemble forecasts where $f_{med} \geq \theta$, while forecasts with ensemble median under the threshold are used to train the light-tailed (TN) one. This technique is applied for calibrating the UWME and the ECMWF ensemble forecasts, see Section 3.3.1. However, in the case of the ALADIN-HUNEPS ensemble (Section 3.3.1) one has only 10 observation stations, so there are not enough data for

separate training of the component models. In such situations one might utilize the same training data set both for the light-tailed and heavy-tailed predictive distribution and then choose between these two models according to the value of the ensemble median. This particular idea is applied in Section 3.3.3 for the ALADIN-HUNEPS forecasts.

We remark that, as an alternative to the use of a fixed threshold over the whole data set, one might also apply an "adaptive" estimation procedure, where for each forecast date the threshold parameter is re-estimated as a fixed quantile of the ensemble medians in the corresponding training period. However, for none of the investigated ensembles and combination models, this adaptive threshold parameter estimation procedure results in significant improvements of the scores and therefore we focus on the computationally simpler procedures using a fixed threshold value.

EMOS models based on combining two parametric families by exclusively selecting one of them at each forecast instance also suffer from the drawback that a suitable covariate has to be chosen as a selection criterion. This necessary step limits the flexibility of the combination models in practice as the adequacy of covariates might depend on the data set at hand. While the ensemble median works reasonably well in the data sets considered here, this observation might change for different EPSs.

3.2.5 Mixture model

In order to combine the advantages of lighter and heavier-tailed distributions flexibly and to avoid the aforementioned problems in the process, we introduce new EMOS models based on weighted mixtures of two parametric distributions.

In particular, we propose to model wind speed with a weighted mixture of models (3.2.4) and (3.2.8) (or (3.2.5) and (3.2.9) for groups of exchangeable ensemble members) resulting in the predictive PDF

$$\psi(x|\,\mu_{TN},\sigma_{TN};\mu_{LN},\sigma_{LN};\omega) := \omega g(x|\,\mu_{TN},\sigma_{TN}) + (1-\omega)h(x|\,\mu_{LN},\sigma_{LN}), \qquad (3.2.13)$$

where the dependence of parameters μ_{TN}, σ_{TN} and μ_{LN}, σ_{LN} on the ensemble are given by (3.2.4) (or (3.2.5)) and (3.2.8) (or (3.2.9)) and (3.2.7), respectively (Baran and Lerch, 2016). In the case of model (3.2.13), location and scale/shape parameters of the TN and LN models, together with the weight $\omega \in [0,1]$, are estimated simultaneously by optimizing some verification score over the training data.

Note that instead of a LN distribution, in (3.2.13) one can incorporate other non-negative laws with heavy right tails. A natural choice would be the generalized Pareto distribution (GPD) used in extreme value theory (see e.g. Bentzien and Friederichs, 2012), however, tests for the ensemble forecasts considered in our case studies indicate a worse predictive performance of the TN-GPD model compared with the TN-LN mixture and the benchmark models.

In comparison with the basic and regime-switching EMOS approaches of Sections 3.2.1 – 3.2.4, the new mixture model exhibits desirable properties from a theoretical perspective as it does not require the exclusive choice of one of multiple parametric families and is more flexible than models based on single parametric distributions. Its advantages from a practical perspective such as a significantly improved calibration will be demonstrated in Section 3.3.3.

3.2.6 Parameter estimation details

All EMOS models specified above have some parameters to be estimated. Similar to Section 2.2, we follow the optimum score principle of Gneiting and Raftery (2007) and minimize the mean value of a proper scoring rule over the training data. The usual candidates are the CRPS and LogS defined by (1.4.2) and (1.4.3), respectively. According to the arguments of Gneiting and Raftery (2007), from these two candidates the use of the mean CRPS in general results in more robust parameter estimation procedures.

For wind speed models of Sections 3.2.1 – 3.2.3 based on single parametric families CRPS can be expressed in closed form allowing efficient optimization procedures. The CRPS of the TN distribution $\mathcal{N}_0^\infty\big(\mu,\sigma^2\big)$ is obviously a special case of (2.2.15) with $a=0$ and $b=\infty$, namely

$$\begin{aligned}\mathrm{CRPS}\Big(\mathcal{N}_0^\infty\big(\mu,\sigma^2\big),x\Big) = \sigma\Bigg[\frac{x-\mu}{\sigma}\Phi\big(\mu/\sigma\big)\Big(2\Phi\big((x-\mu)/\sigma\big)+\Phi\big(\mu/\sigma\big)-2\Big)\\ +2\varphi\big((x-\mu)/\sigma\big)\Phi\big(\mu/\sigma\big)-\frac{1}{\sqrt{\pi}}\Phi\big(\sqrt{2}\mu/\sigma\big)\Bigg]\Big[\Phi\big(\mu/\sigma\big)\Big]^{-2}\end{aligned}$$

for $x\geq 0$, which formula was derived by Thorarinsdottir and Gneiting (2010). Using direct calculations one can also verify that the CRPS of the LN distribution (Baran and Lerch, 2015) is

$$\begin{aligned}\mathrm{CRPS}\Big(\mathcal{LN}\big(\mu,\sigma^2\big),x\Big) = x\Big[2\Phi\big((\log x-\mu)/\sigma\big)-1\Big]\\ -2\mathrm{e}^{\mu+\sigma^2/2}\Big[\Phi\big((\log x-\mu)/\sigma-\sigma\big)+\Phi\big(\sigma/\sqrt{2}\big)-1\Big],\quad x\geq 0.\end{aligned}$$

The closed form of the CRPS for the GEV distribution with non-zero shape ξ equals

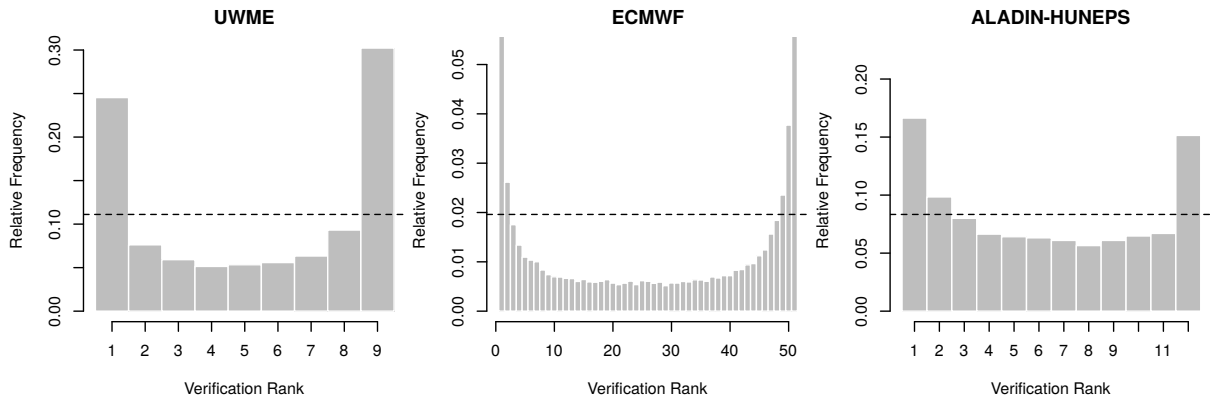
$$\begin{aligned}\mathrm{CRPS}\Big(\mathcal{GEV}\big(\mu,\sigma,\xi\big),x\Big) = \big[\mu-x-\sigma/\xi\big]\big[1-2H(x|\mu,\sigma,\xi)\big]\\ -\frac{\sigma}{\xi}\Big[2^\xi\Gamma(1-\xi)-2\Gamma_\ell\big(1-\xi,-\log H(x|\mu,\sigma,\xi)\big)\Big],\end{aligned}$$

where Γ and Γ_ℓ denote the gamma and the lower incomplete gamma functions, respectively, whereas for $\xi=0$ one has

$$\mathrm{CRPS}\Big(\mathcal{GEV}\big(\mu,\sigma,\xi\big),x\Big) = \mu-x+\sigma[\gamma-\log x]-2\sigma\mathrm{Ei}\big(\log H(x|\mu,\sigma,\xi)\big)$$

with $\gamma\approx 0.5772$ denoting the Euler-Mascheroni constant and $\mathrm{Ei}(x):=\int_{-\infty}^{x}\frac{\mathrm{e}^t}{t}\mathrm{d}t$ being the exponential integral (Friedrichs and Thorarinsdottir, 2012).

In the case studies of Section 3.3.3 for both the TN and the LN EMOS model we estimate model parameters by minimizing the mean CRPS of the predictive distributions and validating observations corresponding to the forecast cases of the training period. However, the GEV model optimization is numerically unstable when using the mean CRPS. Hence in this case, as suggested by Lerch and Thorarinsdottir (2013), parameters are estimated with the help of the ML method optimizing the mean logarithmic score. Objective functions are minimized using the `optim` function in `R` by making use of the popular Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Press *et al.*, 2007, Section 10.9). Obviously, the same approaches to parameter estimation are applied for the TN-LN and TN-GEV regime-switching models.



(a) (b) (c)

Figure 3.1: Verification rank histograms. (a) UWME for the calendar year 2008; (b) ECMWF ensemble for the period 1 May 2010 – 30 April 2011; (c) ALADIN-HUNEPS ensemble for the period 1 April 2012 – 31 March 2013.

In the case of mixture model (3.2.13) the CRPS can be evaluated only numerically, resulting in very long optimization procedures. Hence, we also investigate ML estimation of the parameters and in figures and tables of Section 3.3.3, the corresponding mixture models are denoted by TN-LN mix. (CRPS) and TN-LN mix. (ML). In these cases the Nelder-Mead (Nelder and Mead, 1965) algorithm is applied, which method is slower, but more robust than the BFGS.

3.3 Case studies

The forecast skill of the truncated normal BMA model proposed by Baran (2014) and the EMOS models of (Baran and Lerch, 2015, 2016) are tested in several case studies with the help of UWME, ECMWF and ALADIN-HUNEPS ensemble forecasts of wind speed. Model parameters of all post-processing methods are estimated using the regional approach, described in details in Section 1.3.3.

3.3.1 Data

University of Washington mesoscale ensemble

The 8 members of the UWME are obtained from different runs of the fifth generation Pennsylvania State University–National Center for Atmospheric Research mesoscale model with initial conditions from different sources (Grell *et al.*, 1995). The EPS covers the Pacific Northwest region of western North America providing forecasts on a 12 km grid. Our data base contains ensembles of 48 h forecasts and corresponding validating observations of 10 m maximal wind speed (maximum of the hourly instantaneous wind speeds given in m/s, that is 2-minute averages from the period of two minutes before the hour to on the hour, over the previous 18 hours, see e.g. Sloughter *et al.* (2010)) for 152 stations in the Automated Surface Observing Network (National Weather Service, 1998) in the states of Washington, Oregon, Idaho, California and Nevada in the United States for calendar years 2007 and 2008. The forecasts are initialized at 0 UTC (5 pm local time

when daylight saving time (DST) is in use and 4 pm otherwise) and the generation of the ensemble ensures that its members are not exchangeable. In the present study we investigate only forecasts for calendar year 2008 with additional data from the last month of 2007 used for parameter estimation. Standard quality control procedures were applied to the data set and after removing days and locations with missing data 101 stations remain where the number of days for which forecasts and validating observations are available varies between 160 and 291.

Figure 3.1a shows the verification rank histogram of the raw ensemble, which is strongly U-shaped as in many cases the ensemble members either underepredict or overpredict the validating observations. The ensemble range contains the observed maximal wind speed in only 45.24 % of the cases (the nominal value of this coverage equals 7/9, i.e 77.78 %). Hence, the ensemble is underdispersive and thus uncalibrated and would require statistical post-processing to yield an improved forecast probability density function.

ECMWF ensemble

The global ensemble prediction system of the ECMWF consists of 50 exchangeable ensemble members which are generated from random perturbations in initial conditions and stochastic physics parametrization (Molteni *et al.*, 1996; Leutbecher and Palmer, 2008). Forecasts of near-surface (10 meter) wind speed (given in m/s) for lead times up to 15 days ahead are issued four times a day with a horizontal resolution of about 18 km. Following Lerch and Thorarinsdottir (2013), we focus on the ECMWF ensemble run initialized at 00 UTC (2 am local time when DST operates and 1 am otherwise) and one day ahead forecasts. Predictions of daily maximum wind speed are obtained as the daily maximum of each ensemble member at each grid point location.

The verification is performed over a set of 228 synoptic observation stations over Germany. The validating observations are hourly observations of 10-minute average wind speed measured over the 10 minutes before the hour. Daily maximum wind speed observations are given by the maximum over the 24 hours corresponding to the time frame of the ensemble forecast. Ensemble forecasts at individual station locations are obtained by bilinear interpolation of the gridded model output. Our results are based on a verification period from 1 May 2010 to 30 April 2011, consisting of 83 220 individual forecast cases. Additional data from 1 February 2010 to 30 April 2010 are used to allow for training periods of equal lengths for all days in the verification period and for model selection purposes.

The verification rank histogram of the ECMWF ensemble displayed in Figure 3.1b is even more U-shaped than that of the UWME, and the ensemble range contains the validating observation just in 43.40 % of all cases (here the nominal value is 49/51, that is 96.08 %). Again, the ensemble is underdispersive and statistical calibration is required.

ALADIN-HUNEPS ensemble

The ALADIN-HUNEPS system of the HMS covers a large part of continental Europe with a horizontal resolution of 8 km and is obtained with dynamical downscaling (by the ALADIN limited area model) of the global ARPEGE based PEARP system of Météo France (Horányi *et al.*, 2006; Descamps *et al.*, 2015). The ensemble consists of 11 members, 10 initialized from perturbed initial conditions and one control member from the unperturbed analysis, implying that the ensemble contains groups of exchangeable forecasts.

The data base contains 11-member ensembles of 42 hour forecasts of 10 meter wind speed (given in m/s) for 10 major cities in Hungary (Budapest, Debrecen, Győr, Kecskemét, Miskolc, Nagykanizsa, Nyíregyháza, Pécs, Szeged, Szombathely) produced by the ALADIN-HUNEPS system of the HMS, together with the corresponding validating observations for two different periods. The first data period is between 1 October 2010 and 25 March 2011, whereas the second covers a whole year from 1 April 2012 to 31 March 2013. The validating observations were scrutinized by basic quality control algorithms (e.g. consistency checks) and considered as instantaneous values (valid at a given time), however, they are in fact mean values over the preceding 10 minutes. The model wind speed values are also considered as instantaneous, but they are representatives for a given model time step, which is 5 min in our case.

The forecasts are initialized at 18 UTC (8 pm local time when DST operates and 7 pm otherwise). Both data sets are fairly complete since there are three and six days with missing data. These dates are excluded from the analysis.

Similar to the previous two EPSs, the verification rank histogram of the raw ALADIN-HUNEPS ensemble for the period between 1 April 2012 and 31 March 2013 is far from the desired uniform distribution (see Figure 3.1c), however, it shows a less underdispersive character. The better fit of the ensemble can also be observed on its coverage value of 61.21 %, where the latter should be compared with the nominal coverage of 83.33 % (10/12). The same applies for the verification ranks of the period from 1 October 2010 to 25 March 2011 (not illustrated), here the ensemble coverage is 61.03 %.

3.3.2 BMA modelling of wind speed forecasts

The following two case studies demonstrate the forecast skill of the truncated normal BMA model (3.1.3), which is tested on the ALADIN-HUNEPS ensemble of the HMS and on the 8-member UWME.

Verification results for the ALADIN-HUNEPS ensemble

In the first case study we consider ALADIN-HUNEPS ensemble forecasts of wind speed and the corresponding validating observations for the period between 1 October 2010 and 25 March 2011. This data base coincides with the one investigated in Baran *et al.* (2013), where the authors calibrated the raw ensemble with the help of the gamma BMA model of Sloughter *et al.* (2010) (see Section 3.1.1) considering a training period of 28 calendar days. The optimal training period length was obtained by comparing the mean CRPS of BMA predictive CDFs, the MAE of BMA median and the RMSE of BMA mean forecasts using training periods of $10, 11, \ldots, 60$ days. For details see Baran *et al.* (2013). In this way ensemble members, validating observations and BMA models are available for 146 calendar days (on 20 November 2010 all ensemble members are missing), that is for 1 460 individual forecast cases.

The generation of the ALADIN-HUNEPS ensemble suggests a natural grouping of the members into two exchangeable groups. One contains the control denoted by f_c, whereas in the other are 10 ensemble members corresponding to the different perturbed

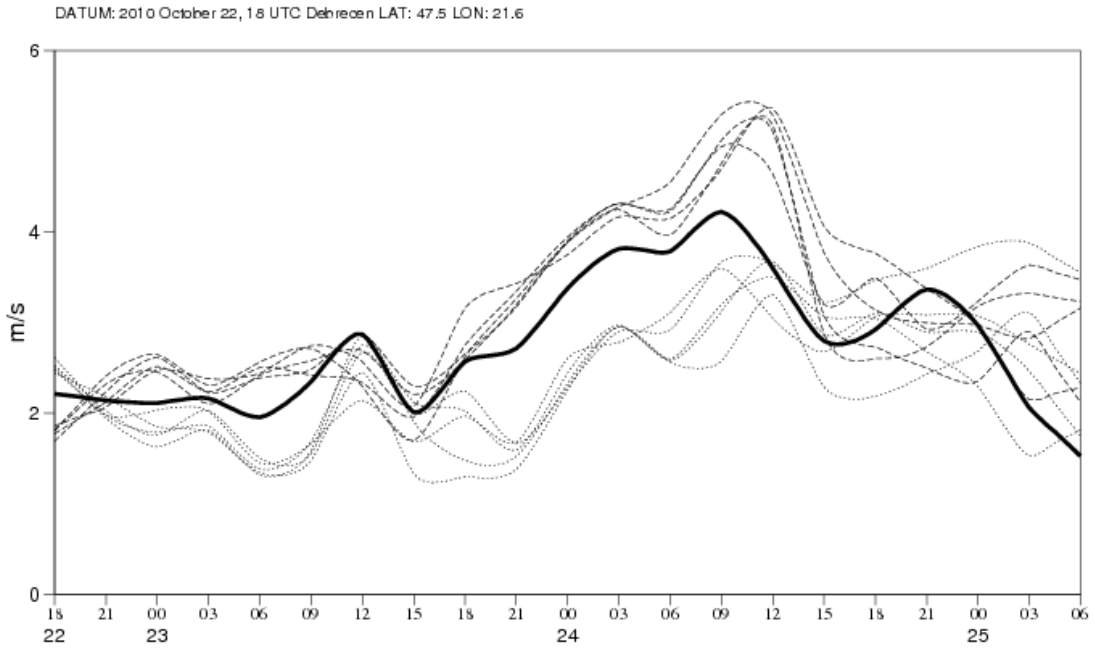


Figure 3.2: Plume diagram of ensemble forecast of 10 m wind speed for Debrecen Airport initialized at 18 UTC, 22 October 2010.

initial conditions denoted by $f_{p,1}, \ldots, f_{p,10}$. This leads us to model

$$p\big(x|\, f_c\, f_{p,1}, \ldots, f_{p,10}; \alpha_c, \alpha_p; \beta_c, \beta_p; \sigma\big) = \omega g_{0,\infty}\big(x|\, \alpha_c + \beta_c f_c, \sigma\big) + \frac{1-\omega}{10} \sum_{\ell=1}^{10} g_{0,\infty}\big(x|\, \alpha_p + \beta_p f_{p,\ell}, \sigma\big), \tag{3.3.14}$$

where $\omega \in [0,1]$, and $g_{0,\infty}$ is defined by (2.1.2).

However, by investigating a bit more the raw ensemble, one can realize that there is a possibility to further distinguish between some of the members (as it was also done for temperature data investigated in Baran *et al.* (2014a)). This is demonstrated by Figure 3.2, where the plume diagram of the ensemble forecasts of 10 m wind speed for Debrecen initialized at 18 UTC, 22 October 2010 can be seen. Figure 3.2 indicates that in practice two clusters of the 10 exchangeable members can be distinguished (particularly look at the 9–15 h and 24–42 h forecast ranges). The two different groups can be linked with the ensemble members created by adding (odd numbered members) and subtracting (even numbered members) 5 perturbations to/from the unperturbed initial conditions (Horányi *et al.*, 2011). Consequently, the behaviour of ensemble member groups $\{f_{\ell,1},\ f_{\ell,3},\ f_{\ell,5},\ f_{\ell,7},\ f_{\ell,9}\}$ and $\{f_{\ell,2},\ f_{\ell,4},\ f_{\ell,6},\ f_{\ell,8},\ f_{\ell,10}\}$ differ from each other.

Therefore, one can also consider a model with three exchangeable groups: control, odd numbered exchangeable members and even numbered exchangeable members. This

Forecast		CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
Two groups	TN-N BMA	0.723	1.063	83.49	3.639
	TN-MC BMA	0.706	1.052	84.32	3.693
	TN-ML BMA	0.707	1.052	84.25	3.674
	Gamma BMA	0.758	1.068	83.56	3.791
Three groups	TN-N BMA	0.721	1.061	83.84	3.627
	TN-MC BMA	0.704	1.048	84.32	3.675
	TN-ML BMA	0.704	1.049	84.18	3.663
	Gamma BMA	0.755	1.064	83.42	3.760
Raw ensemble		0.860	1.122	61.92	2.597

Table 3.1: Mean CRPS of probabilistic, MAE of median forecasts, coverage and average of nominal 83.33 % central prediction interval for the ALADIN-HUNEPS ensemble.

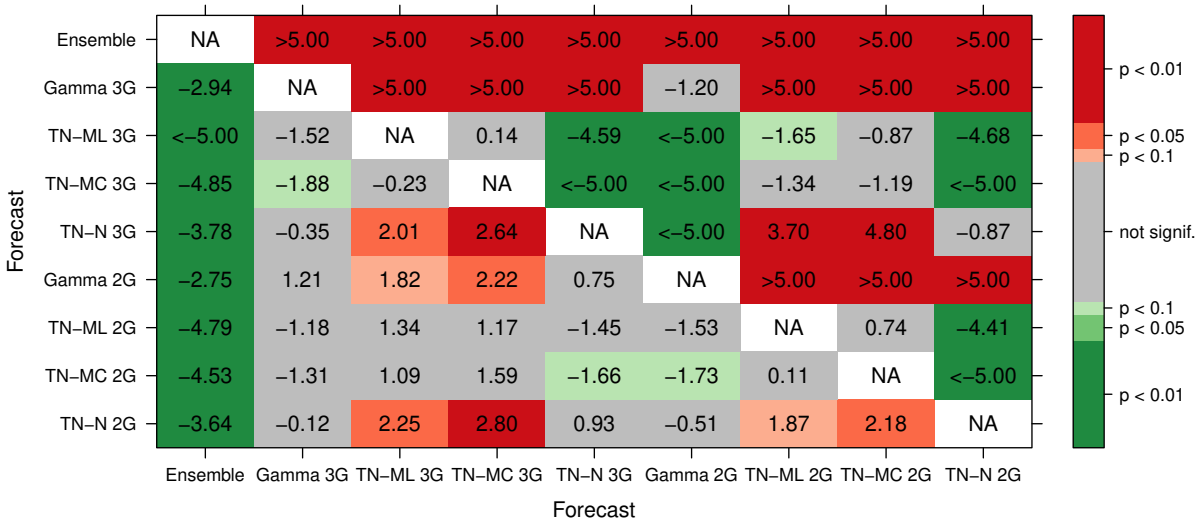


Figure 3.3: Values of the test statistic of the two-tailed DM test for equal predictive performance based on CRPS (*upper triangle*) and absolute error of median forecasts (*lower triangle*) for the ALADIN-HUNEPS data. Green/red entries indicate superior performance of the forecast in the corresponding row/column.

approach results in the predictive PDF

$$q\big(x|\, f_c, f_{p,1}, \ldots, f_{p,10}; \alpha_c, \alpha_o, \alpha_e; \beta_c, \beta_o, \beta_e; \sigma\big) = \omega_c g_{0,\infty}\big(x|\, \alpha_c + f_c\beta_c, \sigma\big) + \sum_{\ell=1}^{5}\Big(\omega_o g_{0,\infty}\big(x|\, \alpha_o + \beta_o f_{p,2\ell-1}, \sigma\big) + \omega_e g_{0,\infty}\big(x|\, \alpha_e + \beta_e f_{p,2\ell}, \sigma\big)\Big), \tag{3.3.15}$$

where for weights $\omega_c, \omega_o, \omega_e \in [0,1]$ we have $\omega_c + 5\omega_o + 5\omega_e = 1$.

In Table 3.1 the mean CRPS of different probabilistic forecasts and MAE of median forecasts are given together with the coverage and average width of nominal 83.33 %

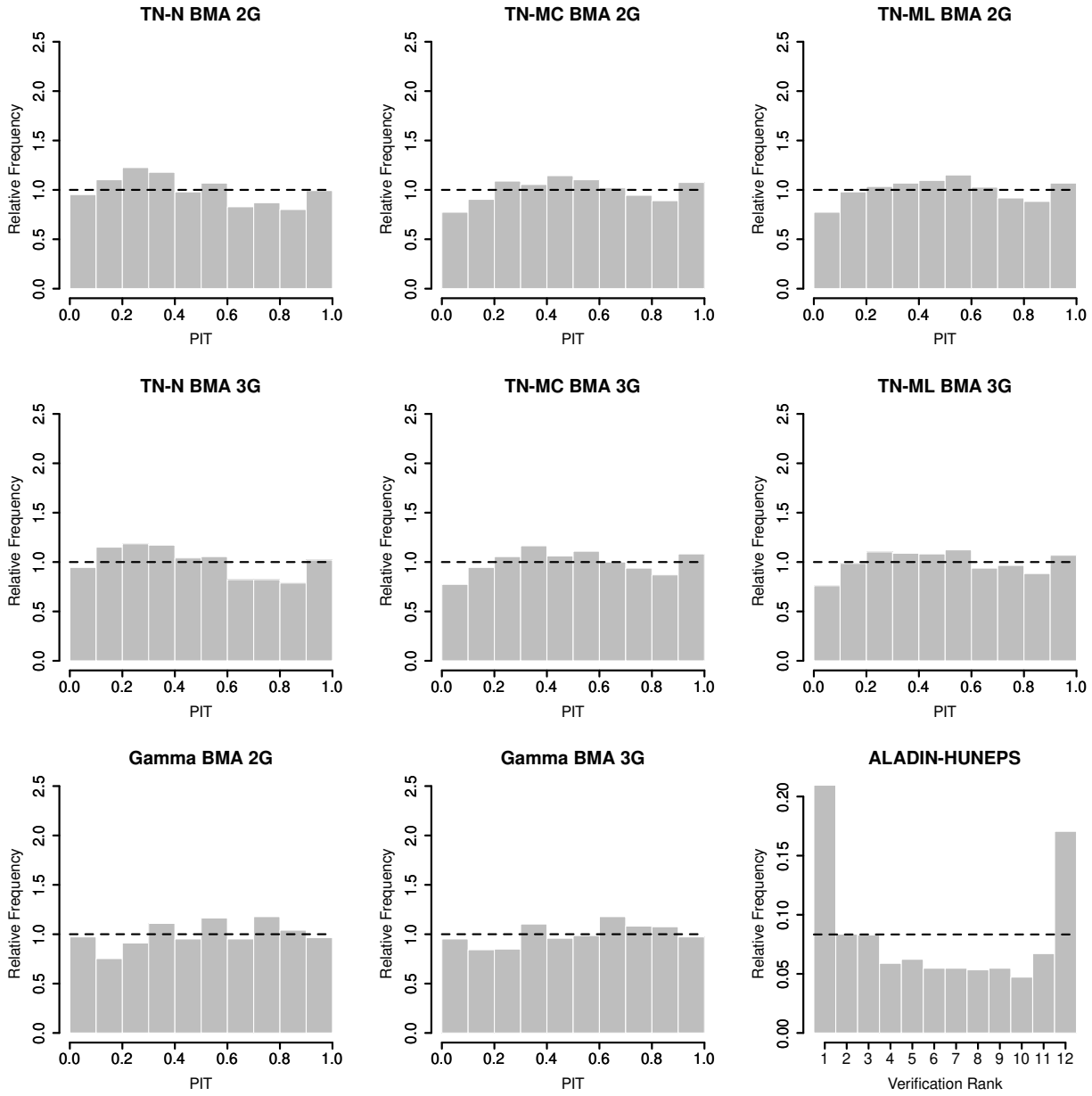


Figure 3.4: Pit histograms of BMA post-processed ALADIN-HUNEPS ensemble forecasts of wind speed and verification rank histogram of the raw ensemble for the period 30 October 2010 – 25 March 2011.

central prediction intervals. Verification measures of probabilistic forecasts and point forecasts calculated using three versions of truncated normal BMA models (3.3.14) and (3.3.15) (TN-N: naive; TN-MC: mean corrected; TN-ML: pure ML) are compared with the corresponding measures calculated for the raw ensemble and applying gamma BMA post-processing (Baran *et al.*, 2013). We remark that gamma BMA models had also been fit to powers 1/3, 1/2, 3/2 and 2 of wind speed (see Sloughter *et al.*, 2010), but the untransformed model gave the best results. Compared with the raw ensemble all BMA post-processed forecasts show a significant decrease in all verification scores considered. The results of the DM tests for equal predictive performance in terms of mean CRPS and

	TN-N BMA	TN-MC BMA	TN-ML BMA	Gamma BMA
Two groups	9.41×10^{-4}	2.11×10^{-3}	4.17×10^{-3}	4.28×10^{-2}
Three groups	3.02×10^{-4}	5.63×10^{-4}	1.67×10^{-2}	7.28×10^{-2}

Table 3.2: Significance levels of α^0_{1234} tests for uniformity of PIT values corresponding to two- and three-group models for the ALADIN-HUNEPS ensemble.

absolute error of median forecasts are given in Figure 3.4. Further, as the listed CRPS and MAE values show, the fit of the truncated normal BMA probabilistic and point forecasts to the validating observations is better than the fit of the gamma BMA ones.

Concerning calibration, one can observe that the coverage values of BMA central prediction intervals are rather close to the correct coverage for all models considered, whereas the coverage values of the central prediction intervals calculated from the raw ensemble are quite poor. This shows that BMA post-processing greatly improves calibration. Further, the truncated normal BMA models yield a bit sharper predictions than the gamma BMA forecasts, and one can also observe that the three-group model slightly outperforms the two-group one.

These results are fully in line with the PIT histograms and verification rank histogram displayed in Figure 3.4. The U-shape of the latter indicates strong underdispersion, which is nicely corrected by post-processing. Instead of the Kolmogorov-Smirnov test applied in the case study of Chapter 2, here we make use of the moment-based α^0_{1234} test proposed by Knüppel (2015) to test uniformity of PIT values. According to the significance levels given in Table 3.2, one can accept uniformity just for the three group gamma BMA model, but the reported p-values are in accordance with the shapes of the histograms and might be used for ranking the different post-processing approaches. Note that Kolmogorov-Smirnov test results in different ranking. Uniformity of PIT values can be accepted for the TN-ML BMA with both groupings of the ensemble members and for the three-group TN-MC BMA, the corresponding p-values are 0.129, 0.176 and 0.056, respectively.

Finally, we remark that similar to Chapter 2, the mean correction step (2.1.10) in the pure ML parameter estimation method seems to be necessary. Running the algorithm without it e.g. for the three group model yields slightly smaller mean CRPS (0.702) but larger MAE value (1.050) and wider central prediction intervals.

Verification results for the University of Washington mesoscale ensemble

As a contrast to the ALADIN-HUNEPS ensemble, the members $f_1, f_2, \ldots, f_8$ of the UWME are non exchangeable. In this way the corresponding BMA model is

$$p\big(x|\, f_1, \ldots, f_8; \alpha_1, \ldots, \alpha_8; \beta_1, \ldots, \beta_8; \sigma\big) = \sum_{\ell=1}^{8} \omega_\ell g\big(x|\, \alpha_\ell + \beta_\ell f_\ell, \sigma\big), \tag{3.3.16}$$

where weights ω_ℓ, $\ell = 1, 2, \ldots, 8$, satisfy $\sum_{\ell=1}^{8} \omega_\ell = 1$. For estimation of model parameters we use the same 25 days training period as in Sloughter *et al.* (2010), where the authors calibrated wind speed forecasts of the UWME for a different time period (1 November 2002 – 31 December 2003). As before, we consider the performance of the BMA predictive PDF (3.3.16) under all three parameter estimation methods of Section

Forecast	CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
TN-N BMA	1.097	1.575	80.45	4.925
TN-MC BMA	1.084	1.560	80.88	4.940
TN-ML BMA	1.077	1.553	81.07	4.957
Gamma BMA	1.112	1.573	78.85	4.828
Raw ensemble	1.353	1.655	45.24	2.532

Table 3.3: Mean CRPS of probabilistic, MAE of median forecasts, coverage and average width of nominal 77.78 % central prediction interval for the UWME.

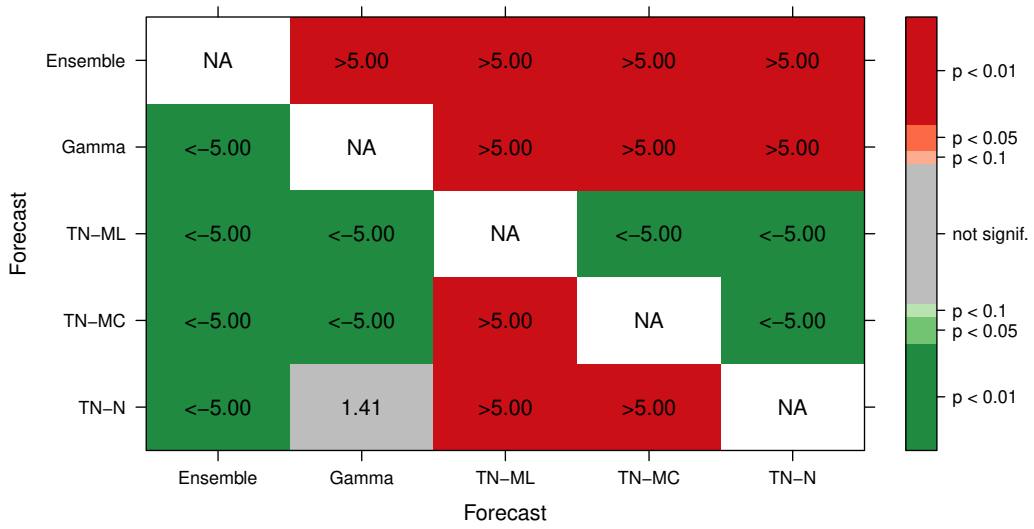


Figure 3.5: Values of the test statistic of the two-tailed DM test for equal predictive performance based on CRPS (*upper triangle*) and absolute error of median forecasts (*lower triangle*) for the UWME data. Green/red entries indicate superior performance of the forecast in the corresponding row/column.

3.1.2, whereas the gamma BMA model of Sloughter *et al.* (2010) is used as a benchmark. Ensemble forecasts for the calendar year 2008 are calibrated, where in total we have 27 481 individual forecast cases.

Table 3.3 shows verification scores of UWME probabilistic and point forecasts and coverage and average width of the nominal 77.78 % central prediction interval, whereas the corresponding results of the DM tests for equal predictive performance in terms of mean CRPS and absolute error of median forecasts are given in Figure 3.4. Compared with the raw ensemble, TN BMA calibration results in coverage values rather close to the correct coverage, significantly lower CRPS and MAE values and wider central prediction intervals. However, the latter fact is a natural consequence of the highly underdispersive character of the raw ensemble. From the three competing parameter estimation methods the pure ML approach yields the lowest CRPS and MAE values combined with the highest coverage and widest central prediction intervals. For the UWME, gamma BMA calibration results in

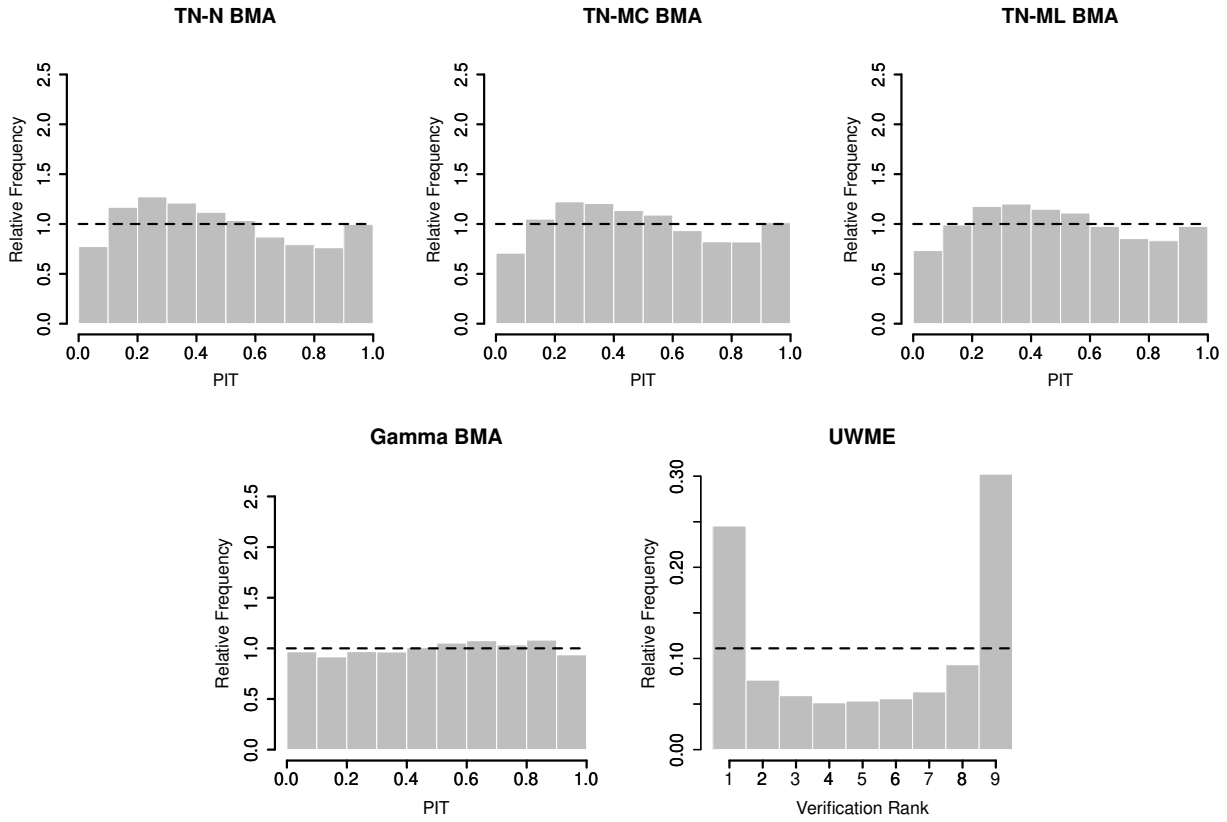


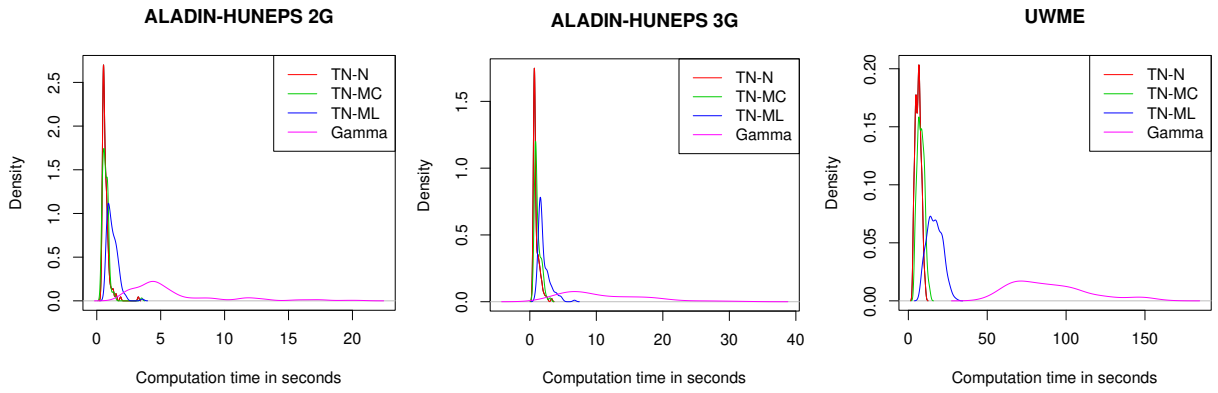
Figure 3.6: Pit histograms of BMA post-processed UWME wind speed forecasts and verification rank histogram of the raw ensemble for calendar year 2008.

Forecast	TN-N BMA	TN-MC BMA	TN-ML BMA	Gamma BMA
Mean p-value	1.83×10^{-12}	1.30×10^{-10}	1.78×10^{-8}	0.088

Table 3.4: p-values of α^0_{1234} tests for uniformity of PIT values for the UWME. Means of 10000 random samples of sizes 2500 each.

significantly higher CRPS and MAE values than the TN BMA except the naive approach to parameter estimation, where there is no difference in the predictive performance of the absolute errors. Further, gamma BMA results in a coverage closer to the nominal value and slightly sharper central prediction intervals, which is also reflected in the PIT histograms of Figure 3.6. All PIT histograms are far closer to the uniform distribution than the verification rank histogram of the raw ensemble, and the slight hump shape of PITs of the three BMA approaches is in line with the larger coverage values of Table 3.3.

Although the α^0_{1234} test rejects the uniformity of PIT values for all post-processing approaches, one can reasonably quantify the differences in calibration by considering the mean p-values of random samples of PITs. Here we take 10000 random samples of sizes 2500 each, and the mean p-values for the different calibration methods given in Table 3.4 are consistent with the shapes of the corresponding histograms of Figure 3.6. Note that a similar ranking of BMA models can be obtained using the Kolmogorov-Smirnov test.



(a) (b) (c)

Figure 3.7: Densities of computation times for the truncated normal and gamma BMA models. ALADIN-HUNEPS ensemble for the period 30 October 2010 – 25 March 2011 with (a) two groups and (b) three groups of exchangeable members; (c) UWME for the calendar year 2008.

Computational aspects of BMA post-processing

For all BMA methods which have been developed so far the most time consuming part of ensemble post-processing is the EM algorithm applied for ML estimation of parameters. For the model presented in Section 3.1.2 we make use of the truncated data EM algorithm for Gaussian mixture models (Lee and Scott, 2012), for the details of parameter estimation see Section 2.1.2 or Baran (2014). In this way, similar to the BMA model with normal components (Raftery *et al.*, 2005), we have closed formulae both in expectation (E) and in maximization (M) steps, whereas in the M step of the gamma BMA model of Sloughter *et al.* (2010) a numerical optimization is used. This difference results in a reasonable gain in speed in favour of the TN BMA approach, which can clearly be observed in Figure 3.7 displaying the kernel density estimates of the distribution of computation times for the various post-processing approaches. Speed tests were made on a portable computer under a 64 bit Fedora 28 operating system (Intel Quad Core i7-4700MQ CPU (2.40GHz $\times$ 4), 20 Gb RAM) using the gamma BMA model of the `ensembleBMA` package of `R` (Fraley *et al.*, 2011) and self-developed codes for the truncated normal BMA approach, which had been adapted to the package. Obviously, from operational point of view the time saved in estimating parameters for a single day is negligible compared with the amount of time needed to create the forecast ensemble. In this way the choice between the two competing methods should be based on their predictive performance. However, if one has to perform modelling for a long time period, moreover, repeating it several times e.g. using different training period lengths in order to determine the optimal one (see e.g. Raftery *et al.*, 2005; Baran *et al.*, 2013, 2014a,b), this small daily difference in speed saves a lot of computation time.

3.3.3 EMOS modelling of wind speed forecasts

To investigate the predictive performance of the EMOS models introduced in Baran and Lerch (2015) and Baran and Lerch (2016) we consider three different case studies. We

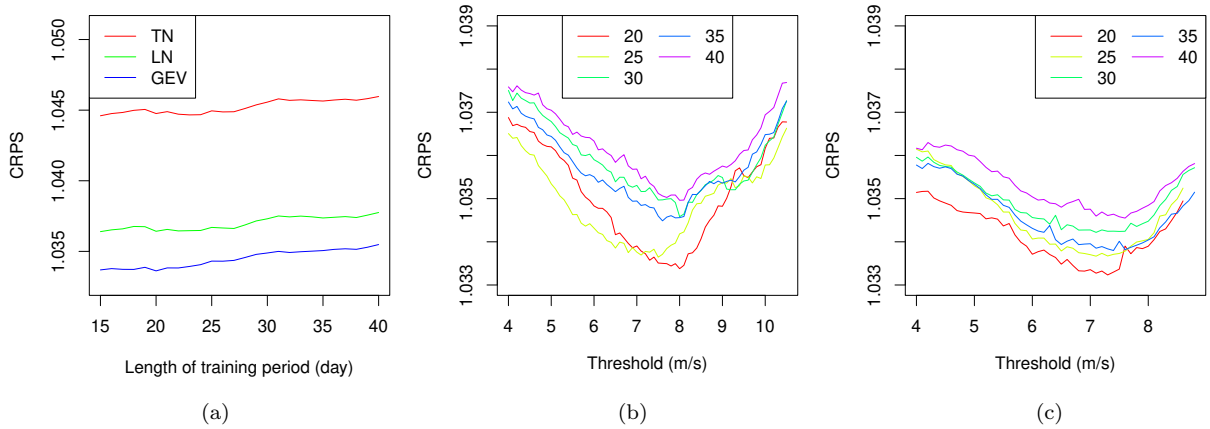


Figure 3.8: Mean CRPS values of the (a) EMOS predictive distributions for various training period lengths; (b) TN-LN mixture models corresponding to different training period lengths as functions of the threshold; (c) TN-GEV mixture models corresponding to different training period lengths as functions of the threshold for the ECMWF ensemble.

compare the forecast skill of the TN, LN and GEV EMOS approaches together with the corresponding regime-switching methods and the TN-LN mixture model with the raw ensemble forecast and climatology. The latter treats the observations from the training period as a forecast ensemble and often used as a baseline for forecast evaluation. We consider wind speed forecasts of the 50-member ECMWF ensemble, the 11-member ALADIN-HUNEPS ensemble of the HMS and the 8-member UWME. As indicated in Section 3.3.1, the three EPSs differ both in generation of the ensemble members and in the predicted wind speed quantity.

Verification results for the ECMWF ensemble

As the fifty members of the ECMWF ensemble are fully exchangeable, the dependencies of the parameters of the TN, LN and GEV distributions on the ensemble members are specified by (3.2.5), (3.2.9) and (3.2.12), respectively, with $K = 1$.

As a first step, we determine the optimal length of the rolling training period valid for all models and the optimal threshold values for TN-LN and TN-GEV regime-switching approaches. Figure 3.8a, showing the mean CRPS values of all three simple EMOS models as functions of the training period length varying from 15 to 40 days, suggests the use of a training period of 20 days. This particular length of the training period is also supported by Figures 3.8b and 3.8c, where mean CRPS values of the TN-LN and TN-GEV regime-switching models, respectively, are plotted as functions of the threshold θ for various training period lengths. According to Figure 3.8b, for the TN-LN model the optimal threshold is 8.0 m/s, whereas for the TN-GEV model similar arguments lead us to a threshold of 5.2 m/s, see Figure 3.8c. Using these parameter values, ensemble forecasts for the one year period between 1 May 2010 and 30 April 2011 are calibrated. In the case of the two regime-switching models an LN distribution is used in around 14 %, while a GEV distribution is applied in about 19 % of the 83 220 individual forecast cases.

Table 3.5 summarizes the verification scores of different probabilistic forecasts together with the average width and coverage of the nominal 96.08 % central prediction intervals.

Forecast		CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
TN-LN mix.	(CRPS)	1.030	1.384	94.34	7.716
TN-LN mix.	(ML)	1.034	1.391	95.81	8.723
TN-LN r.s.	($\theta = 8.0$)	1.033	1.379	92.49	6.363
TN-GEV r.s.	($\theta = 7.3$)	1.033	1.381	92.89	6.600
TN		1.045	1.388	92.19	6.385
LN		1.037	1.386	93.16	6.909
GEV		1.034	1.388	94.84	8.221
Ensemble		1.263	1.441	45.00	1.800
Climatology		1.550	2.144	95.84	11.91

Table 3.5: Mean CRPS of probabilistic, MAE of median forecasts, coverage and average width of 96.08 % central prediction intervals for the ECMWF ensemble.

The improvement with respect to the raw ensemble and climatology is quantified in lower mean CRPS and MAE, and the EMOS predictive PDFs result in calibrated central prediction intervals with coverages very close to the nominal value. The much wider central prediction intervals of the EMOS models compared with the ensemble are a natural consequence of the underdispersive character of the latter.

Among the competing post-processing methods the mixture and regime-switching models clearly outperform the EMOS approaches based on single distributions in almost all scores investigated. The lowest CRPS value belongs to the mixture model with parameters estimated by optimizing the mean CRPS, whereas the regime-switching approaches produce the best MAE scores. The two parameter estimation methods of the TN-LN mixture model make only a very slight difference in model performance (ML estimation leads to slightly worse scores) and the TN-LN mixture EMOS models are able to keep up with the regime-switching approaches.

According to the results of DM tests (not reported), the TN-LN mixture model with ML parameter estimation, the two regime switching models and the GEV EMOS approach do not differ significantly in terms of the mean CRPS. For all other mean CRPS pairs the p-values are less than 0.01 under the null hypothesis of equal predictive performance. Further, in terms of the MAE the only non-significant difference at a 5 % level is between the TN and the GEV EMOS models.

Further, note that the TN, LN and TN-LN regime switching and mixture models are strictly positive, whereas the GEV and TN-GEV models occasionally assign small non-zero probabilities to negative wind speed observations. For the ECMWF data at hand this effect is typically negligible as the average (maximum) probability mass assigned to negative wind speeds is smaller than 0.01 % (5 %) for the GEV model and smaller than 10^{-7} % (0.001 %) for the TN-GEV model.

To assess the predictive ability for high wind speed observations we also compute the twCRPS scores (1.4.8) at threshold values 10, 12 and 15 m/s corresponding approximately to the 90th, 95th and 98th percentiles of the wind speed observations, see Table 3.6. The best scores in the upper tail are obtained by the TN-LN and TN-GEV regime-switching methods followed by the TN-LN mixture model with parameters optimizing the mean CRPS. In almost all cases the relative improvements over the TN model are

Forecast		twCRPS (m/s)		
		$r=10$	$r=12$	$r=15$
TN-LN mix.	(CRPS)	0.194	0.106	0.041
TN-LN mix.	(ML)	0.197	0.108	0.042
TN-LN r.s.	($\theta=8.0$)	0.191	0.103	0.039
TN-GEV r.s.	($\theta=7.3$)	0.191	0.103	0.039
TN		0.200	0.110	0.042
LN		0.198	0.109	0.042
GEV		0.195	0.106	0.041
Ensemble		0.211	0.113	0.043
Climatology		0.251	0.128	0.045

Table 3.6: Mean twCRPS for various thresholds r for the ECMWF ensemble.

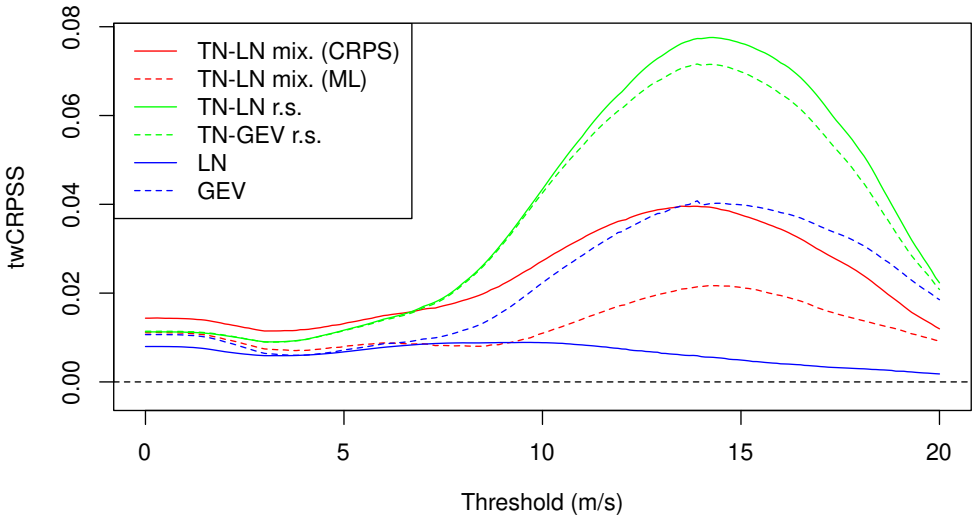


Figure 3.9: twCRPSS values for the ECMWF ensemble with TN as reference model.

considerably higher compared with the improvements in the unweighted CRPS and all score differences in Table 3.6 between the various post-processing methods are significant at a 5 % level. Figure 3.9 further shows the twCRPSS (see (1.4.8)–(1.4.10)) as a function of the threshold employed in the indicator weight function with the TN model as reference forecast. The twCRPSS is strictly positive for all models and threshold values, indicating improvements compared with the TN model. Except for the LN model, the twCRPSS of the various post-processing approaches generally increases for larger threshold values and the greatest relative improvements over the TN model can be detected at threshold values around 15 m/s. Despite the decreasing twCRPSS values of the LN model, the TN-LN regime switching model achieves the largest improvements over the TN EMOS method, closely followed by the TN-GEV regime-switching approach. Further, for large threshold values the GEV EMOS approach is able to catch up with the TN-LN mixture model with parameters estimated by optimizing the mean CRPS.

Figure 3.10 shows the weights ω of the mixture model (3.2.13) estimated using optimizations with respect to the mean CRPS and the mean logarithmic score over the

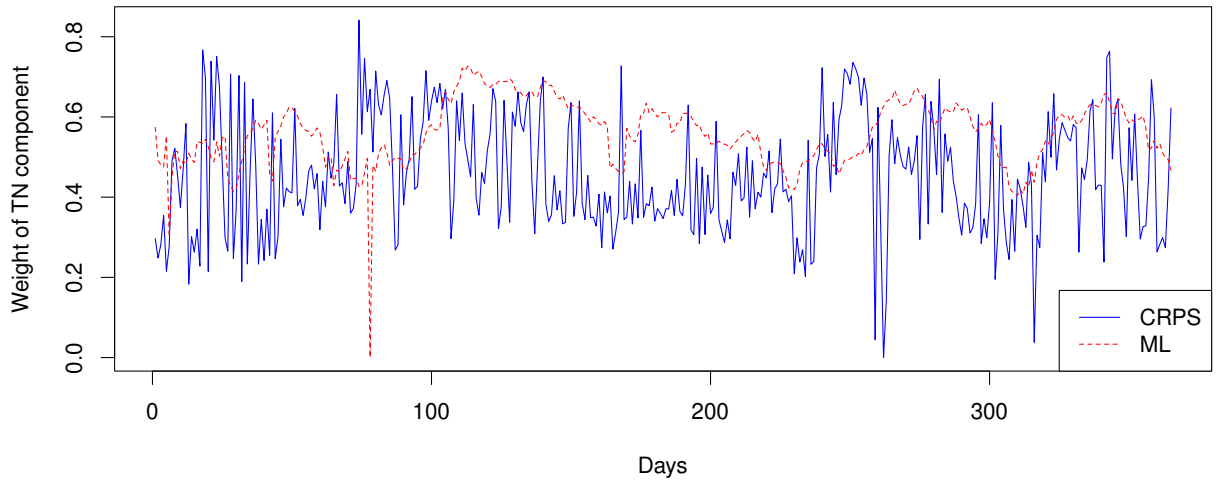


Figure 3.10: Weights of the TN component of the TN-LN mixture model for the ECMWF ensemble.

training data. Despite the similar predictive skills (see Tables 3.5 and 3.6), the two parameter estimating methods result in completely different sets of weights having only a minor non-significant correlation of 0.063. However, having a closer look at the predictive PDFs one can observe that the corresponding locations and scales/shapes of the TN and LN components produced by the two different estimation methods are strongly correlated, their correlations vary between 0.921 and 0.968, except for the scales of the TN component with a correlation of 0.283.

The positive effect of calibration can also be observed on Figure 3.11 showing the PIT histograms of post-processed forecasts and the verification rank histogram of the raw ECMWF ensemble. PIT values of mixture model (3.2.13) with both parameter estimation methods provide the best fit to the desired uniform distribution, whereas the histograms of the two regime-switching approaches are biased, which property is inherited from the TN part of the mixtures. LN EMOS shows similar behaviour to the TN EMOS, however the deviation from uniformity is less pronounced. Finally, as the hump shape of the corresponding PIT histogram indicates, the GEV EMOS model is slightly overdispersed and has too heavy tails.

Similar to Section 3.3.2, we again consider the α^0_{1234} test of uniformity, as it takes into account the dependence of PIT values. By applying it to PITs of all 83 220 individual forecast cases the uniformity is rejected for all models, so further investigations are required. In contrast to the previous case study, now we quantify the differences in calibration by having a look at the rejection rates of the α^0_{1234} test at a 5 % level, based on 10 000 random samples of size 2 500 each, reported in Table 3.7. For the ECMWF data the null hypothesis of uniformity is rejected in all of the cases for all models but the TN-LN mixtures, which is clearly in line with the visual inspection of the PIT histograms in Figure 3.11.

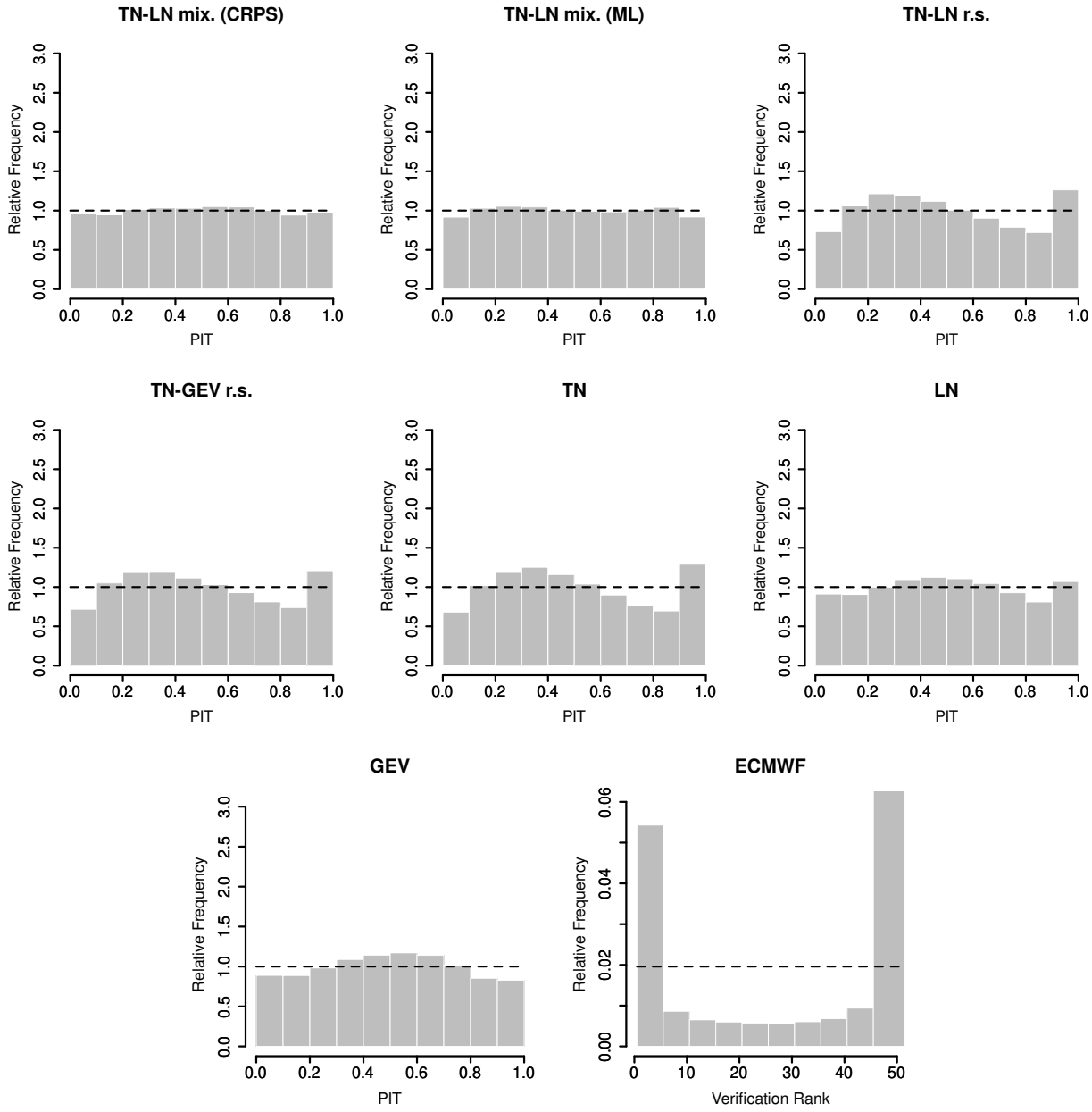


Figure 3.11: Pit histograms of EMOS post-processed ECMWF ensemble forecast of wind speed and verification rank histogram of the raw ensemble for the period 1 May 2010 – 30 April 2011.

Verification results for the ALADIN-HUNEPS ensemble

The data base investigated in this case study differs from the one considered in Section 3.3.2. It contains ensemble forecasts of wind speed and validating observations for the one year period between 1 April 2012 and 31 March 2013 and was first studied from the point of view of statistical calibration in Baran *et al.* (2014b). Similar to the case study of Section 3.3.2, we consider the natural grouping of ensemble members into two groups, where the first group contains just the control member, while in the second are the ten statistically indistinguishable ensemble members initialized from randomly

Forecast		ECMWF	ALADIN-HUNEPS	UWME
TN-LN mix.	(CRPS)	68.72	0	46.85
TN-LN mix.	(ML)	25.58	1.02	30.71
TN-LN r.s.	($\theta = 8.0$)	100.0	100.0	67.22
TN-GEV r.s.	($\theta = 7.3$)	100.0	100.0	95.22
TN		100.0	100.0	100.0
LN		100.0	100.0	100.0
GEV		100.0	78.89	18.49

Table 3.7: Bootstrap estimates of rejection rates (%) of the α^0_{1234} test of uniformity based on 10 000 random samples of size 2 500 each at the 5 % level for the different data sets. Lower rejection rates correspond to better calibrated forecasts with the null hypothesis of uniformity being rejected on fewer occasions.

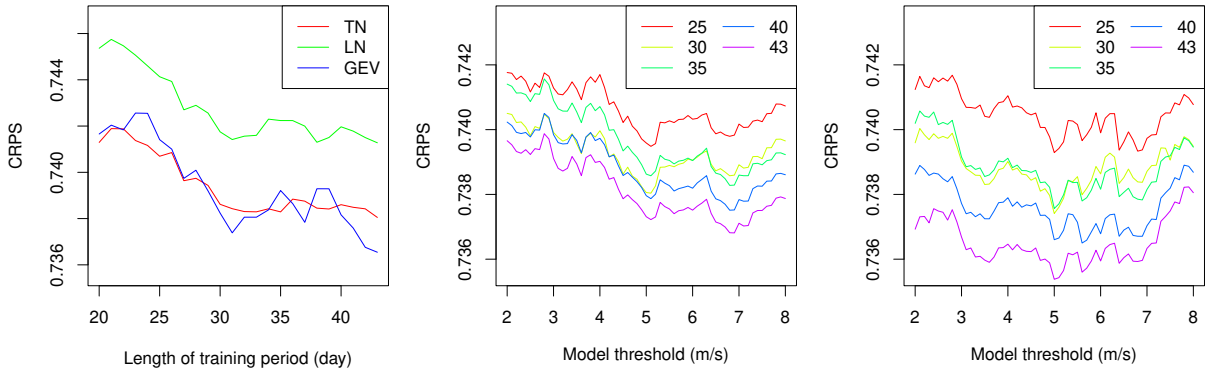


Figure 3.12: Mean CRPS values of the (a) EMOS predictive distributions for various training period lengths; (b) TN-LN mixture models corresponding to different training period lengths as functions of the threshold; (c) TN-GEV mixture models corresponding to different training period lengths as functions of the threshold for the ALADIN-HUNEPS ensemble.

perturbed initial conditions. One should remark here that in Baran *et al.* (2014b) the refined grouping, where the odd and even numbered exchangeable ensemble members form two separate groups, is also studied (see also the three-group model of Section 3.3.2). However, since in the present study the results corresponding to the two- and three-group models are rather similar, only the two-group case is reported.

The detailed study of this particular data set reported in Baran *et al.* (2014a) shows that for the TN distribution based EMOS model, the optimal length of the rolling training period for ALADIN-HUNEPS wind speed forecasts is 43 days. Using this training period length one has a verification period between 15 May 2012 and 31 March 2013 containing 315 calendar days (3 150 forecast cases). Having a look at the mean CRPS values of TN, LN and GEV models as functions of the length of the training period presented in Figure 3.12a, one can derive that this value of 43 days can also be accepted as optimal for all methods, moreover, the optimal TN-LN and TN-GEV thresholds of 6.9 m/s (Figure

Forecast		CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
TN-LN mix.	(CRPS)	0.736	1.037	83.02	3.621
TN-LN mix.	(ML)	0.737	1.040	83.14	3.583
TN-LN r.s.	($\theta=6.9$)	0.737	1.035	83.59	3.535
TN-GEV r.s.	($\theta=5.0$)	0.735	1.039	85.59	3.723
TN		0.738	1.037	83.59	3.534
LN		0.741	1.038	80.44	3.567
GEV		0.737	1.041	81.21	3.541
Ensemble		0.803	1.069	68.22	2.884
Climatology		1.046	1.481	82.54	4.924

Table 3.8: Mean CRPS of probabilistic, MAE of median forecasts and coverage and average width of 83.33 % central prediction intervals for the ALADIN-HUNEPS ensemble.

Forecast		twCRPS (m/s)		
		$r=6$	$r=7$	$r=9$
TN-LN mix.	(CRPS)	0.100	0.053	0.011
TN-LN mix.	(ML)	0.100	0.053	0.011
TN-LN r.s.	($\theta=6.9$)	0.101	0.054	0.011
TN-GEV r.s.	($\theta=5.0$)	0.098	0.052	0.011
TN		0.102	0.054	0.012
LN		0.102	0.054	0.011
GEV		0.098	0.052	0.011
Ensemble		0.112	0.059	0.013
Climatology		0.127	0.064	0.012

Table 3.9: Mean twCRPS for various thresholds r for the ALADIN-HUNEPS ensemble.

3.12b) and 5 m/s (Figure 3.12c), respectively, belong to the 43 days training period, too. The corresponding percentages of usage of LN and GEV distributions in the mixtures are 4 % and 15 %, respectively.

Consider first Table 3.8 reporting the mean CRPS of probabilistic and MAE of median forecasts together with the coverage and average width of the 83.33 % central prediction intervals for the various EMOS models, the ALADIN-HUNEPS ensemble and climatological forecasts. The raw ensemble outperforms climatology and produces sharp forecasts, however, at the cost of being uncalibrated. Post-processing substantially improves the calibration and predictive skill of the raw ensemble. All EMOS models significantly outperform it in terms of mean CRPS and MAE (DM test results are not reported) and have coverages much closer to the nominal value. Regime switching approaches provide the best results, however, the differences in mean CRPS between the mixture and regime switching models are not significant at a 5 % level. The LN EMOS approach results in the highest mean CRPS and significantly underperforms the competitors. In terms of MAE, the TN-LN regime switching model produces the most accurate forecasts, however, the corresponding value does not differ significantly from the MAE of TN and LN EMOS methods and from the mixture model with parameters optimizing the mean CRPS.

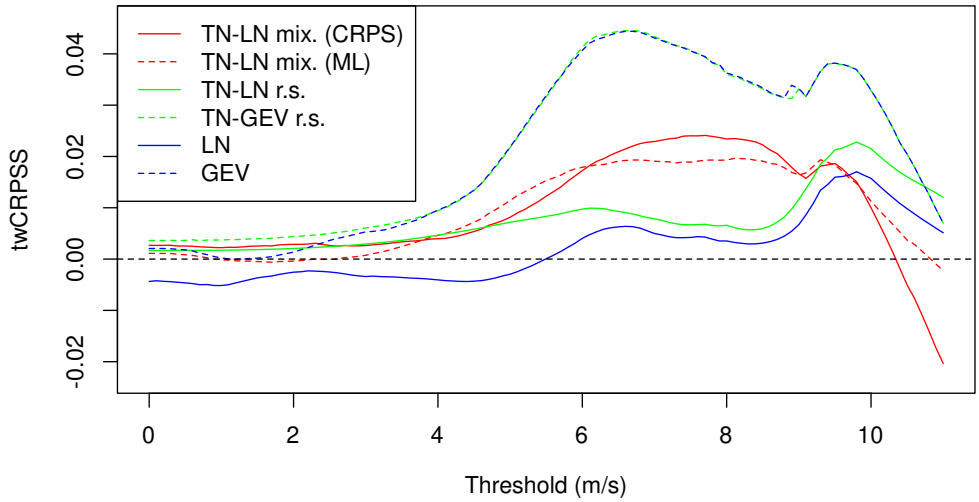


Figure 3.13: twCRPSS values for the ALADIN-HUNEPS ensemble with TN as reference model.

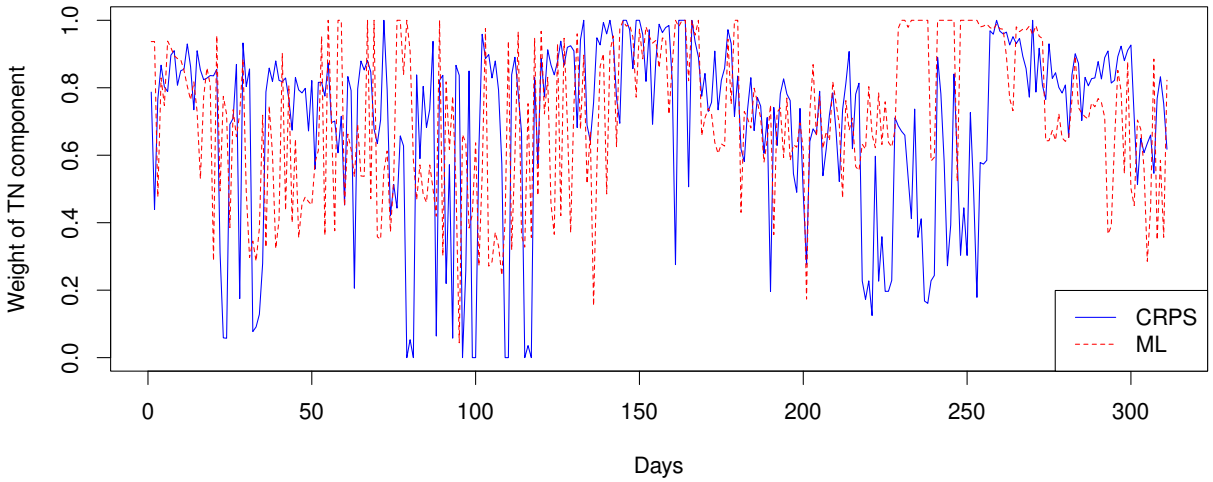


Figure 3.14: Weights of the TN component of the TN-LN mixture model for the ALADIN-HUNEPS ensemble.

Table 3.9 shows the twCRPS scores for three different thresholds corresponding again to the 90th, 95th and 98th percentiles of wind speed observations. For 6 m/s and 7 m/s threshold values the GEV and TN-GEV models result in significantly lower twCRPS scores than all other post-processing approaches (DM test results are not reported), whereas for $r = 9$ m/s there is no statistically significant difference between the competing EMOS models. This phenomenon can also be observed on Figure 3.13, where the twCRPSS values with respect to the reference TN EMOS approach are plotted as functions of the threshold r. Models utilizing GEV distribution provide almost identical curves, have a clear advantage for thresholds between 4 and 10 m/s, but after it the performances decay quickly. However, one should also note that the mean (maximal) probabilities of predicting a negative wind speed by the GEV and TN-GEV regime

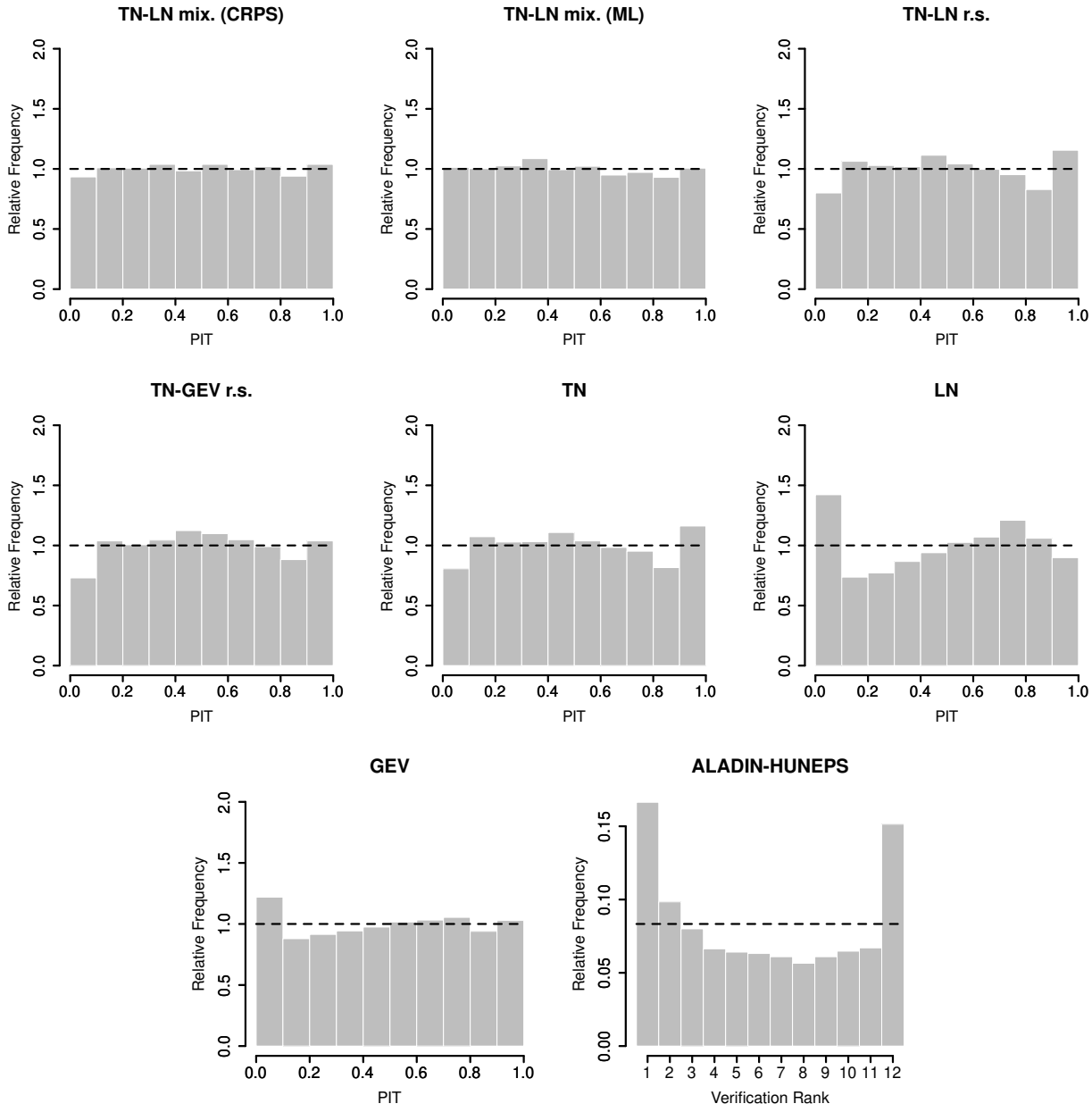
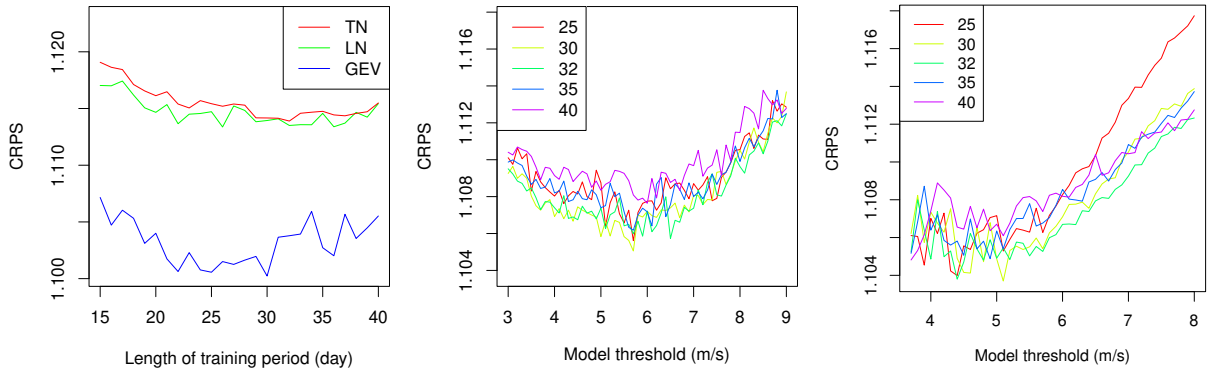


Figure 3.15: Pit histograms of EMOS post-processed ALADIN-HUNEPS ensemble forecast of wind speed and verification rank histogram of the raw ensemble for the period 15 May 2012 – 31 March 2013.

switching methods are 0.33 % (9.46 %) and 2.74×10^{-3} % (0.15 %), respectively.

Further, similar to the previous case study, the weights belonging to the two parameter estimation methods for the TN-LN mixture model (see Figure 3.14) are uncorrelated, whereas the correlations of the corresponding location and scale/shape parameters of the TN (μ_{TN} and σ_{TN}) and LN components (μ_{LN} and σ_{LN}) are 0.875, 0.660 and 0.747, 0.414, respectively.

Finally, let us investigate the PIT histograms of all considered EMOS models, displayed in Figure 3.15. Compared with the verification rank histogram of the raw ensemble, all post-processing methods result in a significant improvement in the goodness of fit to



(a) (b) (c)

Figure 3.16: Mean CRPS values of the (a) EMOS predictive distributions for various training period lengths; (b) TN-LN mixture models corresponding to different training period lengths as functions of the threshold; (c) TN-GEV mixture models corresponding to different training period lengths as functions of the threshold for the UWME.

the uniform distribution, while from the competing calibration methods the TN-LN mixture models have the best performance. These two histograms show no visible tendency in deviation from uniformity, LN and GEV EMOS models result in slightly overdispersed PIT histograms, whereas the histograms of the regime switching approaches mimic the histogram of the TN EMOS model. These shapes are fully in line with the corresponding rejection rates of the α^0_{1234} test reported in Table 3.7.

Verification results for the UWME

The members of the UWME are clearly distinguishable, as they are generated using initial conditions from eight different sources. Hence, location and scale/shape parameters of the TN, LN and GEV distributions are linked to the ensemble via (3.2.4), (3.2.8) and (3.2.11), respectively, with $K = 8$.

To determine the optimal length of the training period for all models and the optimal model thresholds for the regime-switching approaches, we proceed as for the ECMWF and ALADIN-HUNEPS ensemble and compute the mean CRPS over a range of lengths of training periods and choices for the model threshold θ. The mean CRPS of the GEV model takes its minimum at day 30 (see Figure 3.16a) and this training period length seems reasonable for the other two models, too. The use of a 30 day training period is also supported by Figures 3.16b and 3.16c suggesting model thresholds of $\theta = 8.0$ m/s and $\theta = 7.3$ m/s for the TN-LN and TN-GEV regime-switching models, respectively. However, for the UWME the threshold values have much bigger effect than in the case of the ALADIN-HUNEPS ensemble, as the curves corresponding to different training period lengths are very close to each other and often intersect. Using a 30 day training period and the above thresholds, ensemble forecasts for the calendar year 2008 are calibrated. In the case of the two regime-switching models an LN distribution is used in around one third, while a GEV distribution is applied in about 40 % of the 27 481 individual forecast cases. Note that in the case study of Section 3.3.2 a different training period length of 25 days is applied.

Forecast		CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
TN-LN mix.	(CRPS)	1.104	1.551	79.02	4.786
TN-LN mix.	(ML)	1.108	1.560	78.12	4.779
TN-LN r.s.	($\theta = 5.7$)	1.105	1.550	77.73	4.642
TN-GEV r.s	($\theta = 5.2$)	1.105	1.555	77.20	4.597
TN		1.114	1.550	78.65	4.666
LN		1.114	1.554	77.29	4.692
GEV		1.100	1.554	77.20	4.686
Ensemble		1.353	1.655	45.24	2.532
Climatology		1.412	1.987	81.10	5.898

Table 3.10: Mean CRPS of probabilistic, MAE of median forecasts and coverage and average width of 77.78 % central prediction intervals for the UWME.

Forecast		twCRPS (m/s)		
		$r = 9$	$r = 10.5$	$r = 14$
TN-LN mix.	(CRPS)	0.147	0.073	0.010
TN-LN mix.	(ML)	0.147	0.073	0.010
TN-LN r.s.	($\theta = 5.7$)	0.149	0.073	0.010
TN-GEV r.s	($\theta = 5.2$)	0.145	0.072	0.010
TN		0.150	0.074	0.010
LN		0.147	0.073	0.010
GEV		0.145	0.072	0.010
Ensemble		0.175	0.085	0.011
Climatology		0.173	0.081	0.010

Table 3.11: Mean twCRPS for various thresholds r for the UWME.

Mean CRPS of probabilistic, MAE of median forecasts and the coverage and average width of 77.78 % central prediction intervals are reported in Table 3.10. Compared with the raw ensemble and climatology, post-processed forecasts exhibit the same behaviour as before: improved predictive skills and better calibration. The GEV EMOS method provides the smallest mean CRPS, there is no significant difference between the two regime-switching approaches and the mixture model optimizing the mean CRPS, and all post-processing approaches but the LN model significantly outperform the TN EMOS (DM test results are not reported). However, the latter method is the more accurate in terms of the MAE of median forecasts together with the TN-LN regime-switching model and the TN-LN mixture with CRPS optimization. The differences between MAE values of these approaches are again not significant. Note that the 77.73 % coverage of the TN-LN regime-switching model is very close to the nominal coverage of 77.78 % of the raw ensemble, combined with the second sharpest central prediction interval from the seven post-processing approaches. From the two different parameter estimation methods for the TN-LN mixture model, similar to the previous two case studies, the one using the ML approach results in slightly worse predictive performance.

To investigate the forecast skill of the different post-processing approaches for high

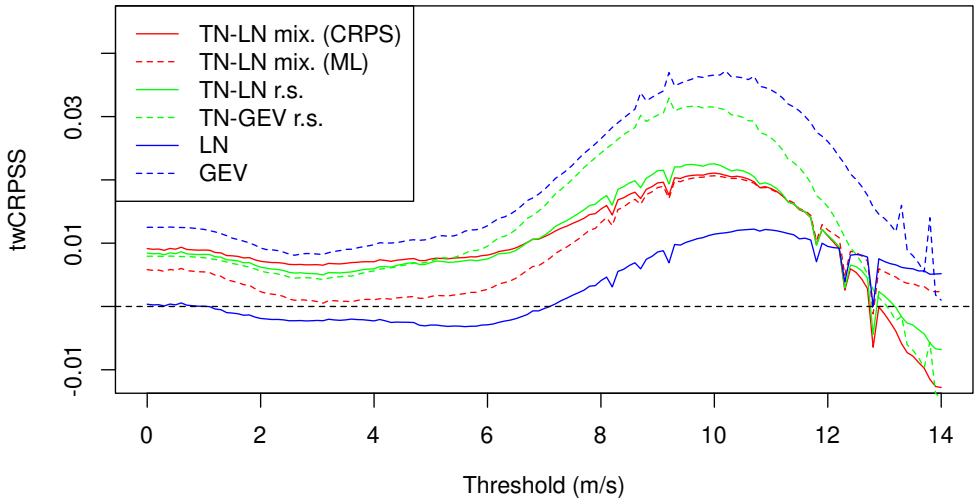


Figure 3.17: twCRPSS values for the UWME with TN as reference model.

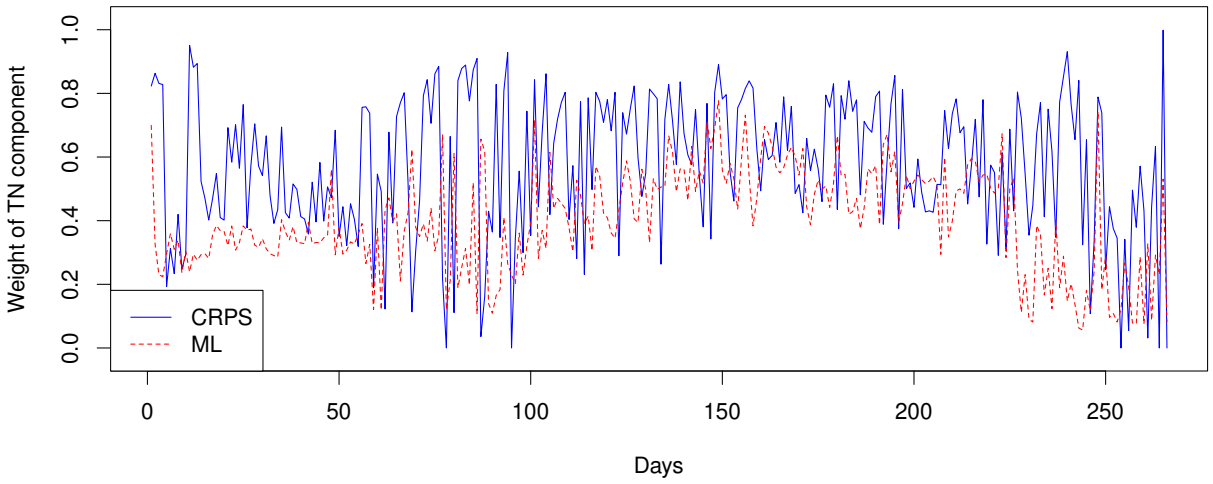


Figure 3.18: Weights of the TN component of the TN-LN mixture model for the UWME.

wind speed, consider first Table 3.11 summarizing the mean twCRPS over calendar year 2008 for three different thresholds. Threshold values 9, 10.5 and 14 m/s again correspond to the 90th, 95th and 98th percentiles of the observed wind speed. Similar to the previous two case studies, GEV and TN-GEV regime-switching EMOS models provide the smallest score and their superiority can also be observed on Figure 3.17 displaying twCRPSS with respect to the TN EMOS as function of the threshold. Note that the advantage of models using GEV distribution is far less pronounced than e.g. in the case of the ECMWF data, but here the GEV EMOS model outperforms its regime switching counterpart. However, we remark that for the GEV model the mean (maximal) probability of forecasting a negative wind speed is around 0.05 % (3.89 %), whereas for the TN-GEV regime switching approach this probability equals 0.02 % (2.67 %).

In contrast to the ECMWF and ALADIN-HUNEPS ensembles, the weights of the TN component of the two versions of model (3.2.13) plotted in Figure 3.18 show a positive

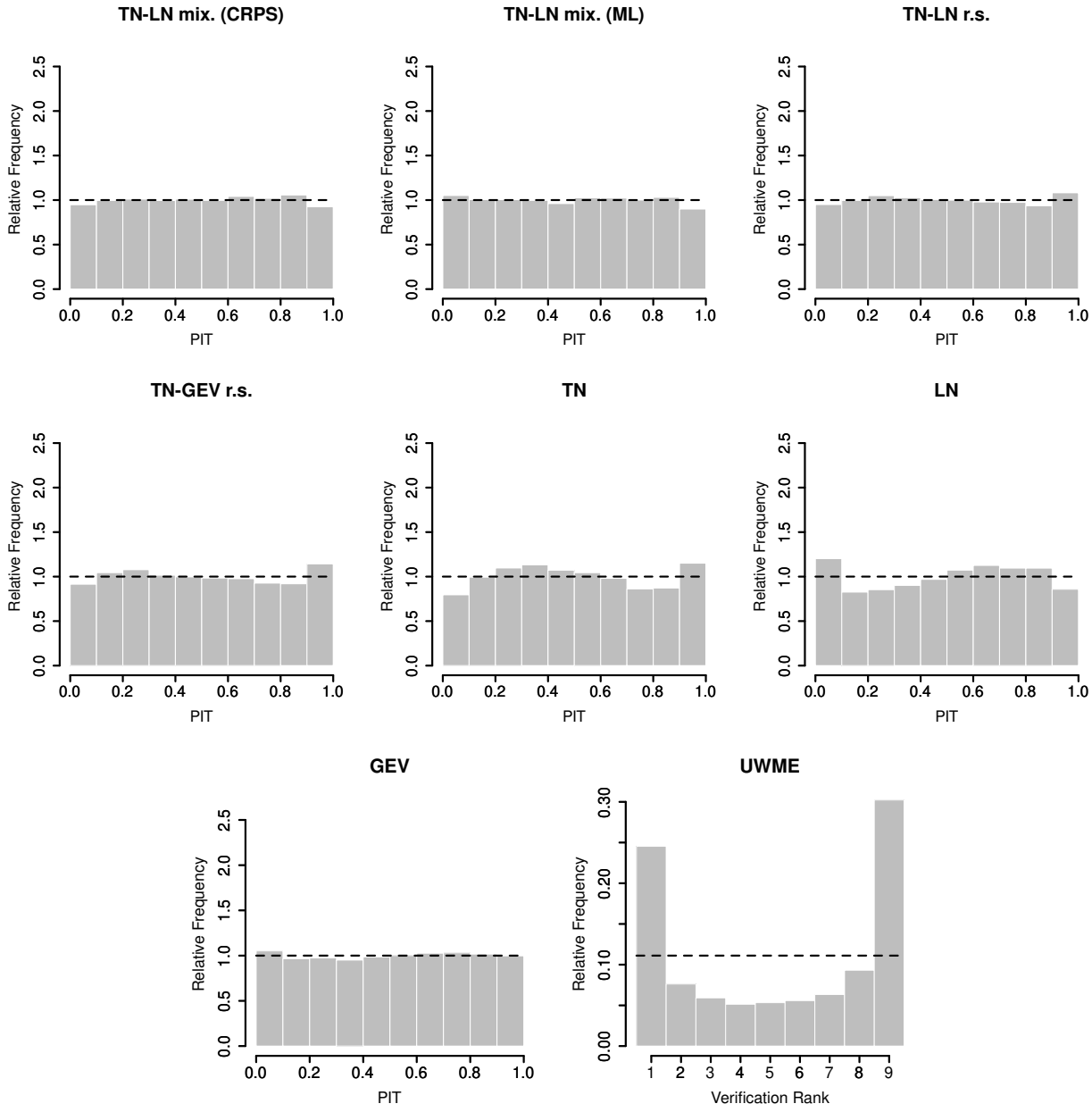
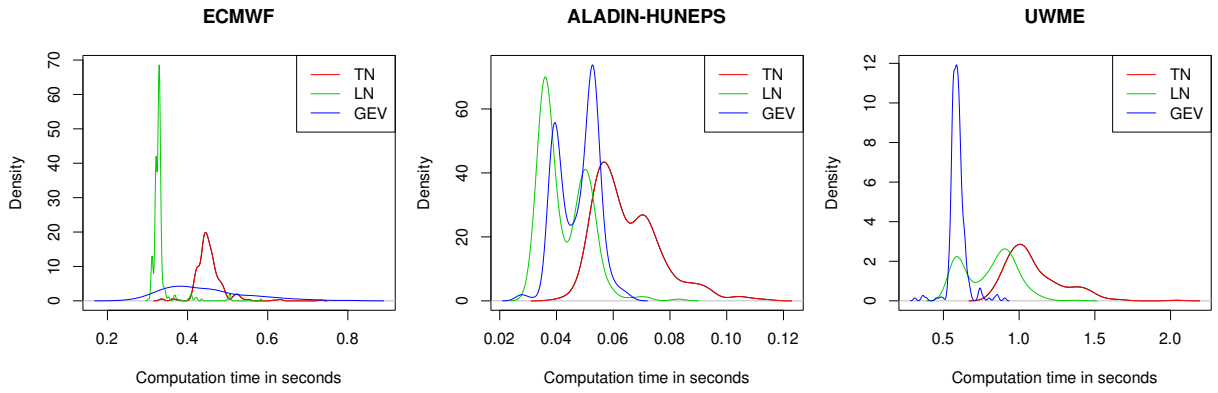


Figure 3.19: Pit histograms of EMOS post-processed UWME forecast of wind speed and verification rank histogram of the raw ensemble.

correlation of 0.214. Further, for the UWME the parameter estimates of μ_{LN} and σ_{LN} exhibit stronger correlations than the estimated location and scale parameters μ_{TN} and σ_{TN} of the TN component, the corresponding values are 0.858, 0.826 and 0.427, 0.259, respectively.

Finally, to get an overview about calibration, consider again the PIT histograms of the EMOS predictive distributions displayed in Figure 3.19. All post-processing approaches are able to significantly improve the underdispersive character of the UWME. The PIT histograms, in general, are much closer to the uniform distribution than in the previous two case studies, just TN and LN EMOS models show a slight overdispersion, which is nicely corrected by combining them using regime-switching. The flattest PIT histograms



(a) (b) (c)

Figure 3.20: Densities of computation times for the TN, LN and GEV models. (a) ECMWF ensemble for the period 1 May 2010 – 30 April 2011; (b) UWME for the calendar year 2008; (c) ALADIN-HUNEPS ensemble for the period 15 May 2012 – 31 March 2013.

correspond to the GEV and mixture models, which is completely in accordance with the results of the goodness of fit test provided in Table 3.7. Note that this is the only situation, when any post-processing approach results in better calibration than the TN-LN mixture models.

Computational aspects of EMOS post-processing

In EMOS post-processing the most computation intensive part is the numerical optimization used in parameter estimation. Figure 3.20a–c show the kernel density estimates of the distribution of computation times over the days in the verification period for the individual EMOS models for the UWME, ECMWF ensemble and ALADIN-HUNEPS ensemble, respectively, calculated on the portable computer specified in Section 3.3.3 (64 bit Fedora 28 operating system, Intel Quad Core i7-4700MQ CPU (2.40GHz $\times$ 4), 20 Gb RAM). In contrast to the previous case study, model parameters were calculated without the help of the general `ensembleMOS` package of `R` (Yuen *et al.*, 2018), using individual codes tailored to the particular tasks. The densities displayed in Figure 3.20 clearly show that in terms of computation time the LN model outperforms both the TN and the GEV method. In the case of the regime-switching approaches, roughly one has to add the computation costs of the component models. We do not present the results for the mixture models as they are not directly comparable. From the one hand they result in larger dimensional optimization problems to be solved using a different optimization algorithm (Nelder-Mead instead of BFGS), from the other hand optimization with respect to the mean CRPS requires a numerical integration at each iteration step when the CRPS is evaluated. Hence, the computation cost of the mixture model using CRPS optimization is about 500 times higher than that of the one using the ML approach. In this way, as the forecast skill of the two TN-LN mixture models is rather similar, the latter is preferred. Finally, one should remark again that the computation cost of all presented post-processing approaches but the above mentioned problematic mixture model is negligible compared with the costs of producing the forecast ensemble, thus the choice of the calibration method should be based merely on the forecast skill.

3.4 Conclusions

In this chapter we investigate various parametric calibration method for ensemble forecasts of wind speed. First we describe a new BMA model for providing a predictive PDF which is a mixture of normal distributions truncated to the left at zero. The model presented here is a special case of the one given in Chapter 2, where the statistical calibration of hydrological ensemble forecasts is investigated. The truncated normal BMA model with three different approaches to parameter estimation is tested on ensemble forecasts of wind speed produced by ALADIN-HUNEPS EPS of the HMS and on the 8-member UWME. Using appropriate verification measures (CRPS of probabilistic, MAE of median forecasts, coverage and average width of central prediction intervals corresponding to the nominal coverage) and graphical tools, the predictive performance of this novel approach is compared with that of the gamma BMA model. Based on the results of these two case studies we conclude that truncated normal BMA post-processing of ensemble predictions of wind speed significantly improves the calibration of probabilistic and accuracy of point forecasts. Further, the predictive performance of the truncated normal BMA model is significantly better than the forecast skill of the gamma BMA method, moreover, in terms of computation time the new approach is more efficient.

Then we study a novel EMOS model for calibrating ensemble forecasts of wind speed providing a predictive PDF which follows a log-normal distribution. In order to have better forecasts in the tails, we also consider a regime-switching approach based on the ensemble median, which considers a truncated normal EMOS model for low values and a log-normal EMOS for the high ones. Even more flexibility can be reached by the use of the mixture predictive PDF modelling wind speed as a weighted mixture of a truncated normal and a log-normal distribution with location and scale/shape parameters depending on the ensemble. Model parameters and mixture weight are estimated simultaneously by optimizing either the mean continuous ranked probabilistic score or the mean logarithmic score (ML estimation) of the predictive distribution over the training data. Similar to the BMA models, the LN EMOS, TN-LN regime-switching EMOS and TN-LN mixture EMOS approaches are tested on wind speed forecasts of the UWME and the ALADIN-HUNEPS ensemble, but now the set of case studies is extended with the calibration of the 50-member ECMWF ensemble forecasts. These EPSs differ both in the wind speed quantities being forecast and in the generation of the ensemble members. The predictive skills of the new model are compared with those of the TN based EMOS method, the GEV and the TN-GEV regime-switching EMOS models, the raw ensemble and the climatological forecasts. In order to assess forecast skill at high wind speed values, besides the verification scores and graphical tools of the other case study of this section, mean twCRPS values corresponding to 90th, 95th and 98th percentiles of the verifying observations are also considered. Compared with the raw ensemble and climatology, the advantage of EMOS post-processing is unquestionable. Considering just EMOS models based on single parametric distributions, the GEV EMOS approach, especially for high wind speed values, results in slightly better calibrated forecasts than the TN and LN EMOS models. However, the GEV EMOS might occasionally predict negative wind speed values. In the case of the ALADIN-HUNEPS ensemble the maximal probability of this event is almost 10 %, which is far beyond being acceptable. A possible solution is to investigate the use of a truncated GEV distribution in EMOS modelling, which is a reasonable direction of future research. TN-LN and TN-GEV regime-switching approaches successfully combine

the advantages of light- and heavy-tailed distributions and result in significant improvements in calibration. Moreover, for the latter model the maximal probability of predicting negative wind speed is also substantially reduced compared with the GEV EMOS. The main difficulty of the regime-switching approaches is in finding an appropriate covariate to select between the heavy- and light-tailed component and in specifying the corresponding threshold. This problem can be solved by the use of e.g. the TN-LN mixture model, which does not require the exclusive choice of one of the parametric families as forecast distribution and it is not necessary to determine suitable covariates for model selection or to estimate model selection threshold over the training period. Obviously, simultaneous estimation of the mixing weight and model parameters of both components is computationally more demanding than the aforementioned EMOS approaches. In our three case studies the mixture model results in the best calibration, however, the small increase in verification scores compared with the regime-switching approaches is often non-significant. Naturally, there are many other ways of combining different post-processing methods in order to improve predictive performance. A detailed comparison of the state of the art techniques including the TN-LN mixture model can be found in Baran and Lerch (2018).

Chapter 4

Models for probabilistic quantitative precipitation forecasting

Statistical calibration of ensemble forecasts of precipitation is far more difficult than the post-processing of e.g. temperature or wind speed. As pointed out by Scheuerer and Hamill (2015), precipitation has a discrete-continuous nature with a positive probability of being zero, and larger expected precipitation amount results in larger forecast uncertainty. Sloughter *et al.* (2007) introduced a BMA model where each individual predictive PDF consists of a discrete component at zero and a gamma distribution modelling the case of positive precipitation amounts. Wilks (2009) uses extended logistic regression to provide full probability distribution forecasts, whereas in EMOS modelling a popular choice is to consider a continuous distribution that can take both positive and negative values and left censor it at zero (Scheuerer, 2014; Scheuerer and Hamill, 2015), thereby assigning the mass of negative values to a zero precipitation accumulation. In this chapter, based on Baran and Nemoda (2016), we introduce a new EMOS model for calibrating ensemble forecasts of precipitation and overview some existing approaches to post-processing providing full predictive distribution of this weather quantity. The forecast skill of the new model is investigated in two case studies dealing with calibration of UWME and ALADIN-HUNEPS ensemble precipitation forecasts.

4.1 Discrete-continuous gamma BMA model

The BMA method suggested by Sloughter *et al.* (2007) treats separately the cases of zero and positive precipitation and provides a predictive distribution of the cube root of precipitation accumulation. Given an ensemble member f_k, the probability of zero precipitation is specified by logistic regression

$$\operatorname{logit} \mathsf{P}(x = 0 |\, f_k) := \log \frac{\mathsf{P}(x = 0 |\, f_k)}{\mathsf{P}(x > 0 |\, f_k)} = a_{0k} + a_{1k} f_k^{1/3} + a_{2k} \mathbb{I}_{\{f_k = 0\}}, \tag{4.1.1}$$

whereas the conditional distribution of the cube root x of precipitation accumulation given that it is positive follows a gamma law $\Gamma(\kappa_k, \theta_k)$ with shape $\kappa_k > 0$ and scale $\theta_k > 0$ with PDF (3.1.1). Mean $\mu_k = \kappa_k \theta_k$ and variance $\sigma_k^2 = \kappa_k \theta_k^2$ are linked to the ensemble members via equations

$$\mu_k = b_{0k} + b_{1k} f_k^{1/3} \qquad \text{and} \qquad \sigma_k^2 = c_0 + c_1 f_k. \tag{4.1.2}$$

The BMA predictive PDF is

$$p(x|\, f_1, \ldots, f_K) = \sum_{k=1}^{K} \omega_k \Big(\mathsf{P}(x = 0|\, f_k) \mathbb{I}_{\{x=0\}} + \mathsf{P}(x > 0|\, f_k) g_k(x|\, f_k) \mathbb{I}_{\{x>0\}} \Big),$$

where $\mathsf{P}(x = 0|\, f_k)$ is defined by (4.1.1) and $g_k(x|\, f_k)$ denotes the gamma PDF with mean and variance specified by (4.1.2).

Parameters a_{0k}, a_{1k}, a_{2k} are estimated from the training data by logistic regression, mean parameters b_{0k}, b_{1k} are obtained using linear regression connecting the cube roots of non-zero observations of precipitation accumulation to the cube roots of the corresponding ensemble members, whereas for estimating weights ω_k and variance parameters c_0, c_1 one uses the maximum likelihood approach with EM algorithm to maximize the likelihood function. For mode details see Sloughter *et al.* (2007).

4.2 EMOS models for precipitation forecasting

As mentioned earlier, EMOS approach provides a single model for the probability of zero precipitation and for the PDF of the positive precipitation amount. Here we present two different methods which are both included in the `ensembleMOS` package of `R` (Yuen *et al.*, 2018), for a detailed overview of the state of the art approaches see Wilks (2018).

4.2.1 Censored and shifted gamma EMOS model

Consider a gamma distribution $\Gamma(\kappa, \theta)$ with shape $\kappa > 0$ and scale $\theta > 0$, specified by the PDF (3.1.1), let $\delta > 0$, and denote by $G(x|\, \kappa, \theta)$ the CDF of the $\Gamma(\kappa, \theta)$ distribution. Then the shifted gamma distribution left censored at zero (CSG) $\Gamma^0(\kappa, \theta, \delta)$ with shape κ, scale θ and shift δ can be defined with CDF

$$G_0(x|\, \kappa, \theta, \delta) := \begin{cases} G(x + \delta|\, \kappa, \theta), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{4.2.3}$$

This distribution assigns mass $G(\delta|\, \kappa, \theta)$ to the origin and has generalized PDF

$$g_0(x|\, \kappa, \theta, \delta) := \mathbb{I}_{\{x=0\}} G(\delta|\, \kappa, \theta) + \mathbb{I}_{\{x>0\}} g(x + \delta|\, k, \theta).$$

Short calculation shows that the mean μ_0 of $\Gamma_0(\kappa, \theta, \delta)$ equals

$$\mu_0 = \theta\kappa\big(1 - G(\delta|\, \kappa + 1, \theta)\big) - \delta\big(1 - G(\delta|\, \kappa, \theta)\big),$$

whereas the p-quantile q_p $(0 < p < 1)$ of (4.2.3) equals 0 if $p \leq G(\delta|\, \kappa, \theta)$, and the solution of $G(q_p + \delta|\, \kappa, \theta) = p$, otherwise.

In the CSG EMOS model proposed by Baran and Nemoda (2016), the ensemble members are linked to the mean μ and variance σ^2 of the underlying gamma distribution via equations

$$\mu = a_0 + a_1 f_1 + \cdots + a_K f_K \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 \overline{f}, \tag{4.2.4}$$

where $\overline{f}$ denotes the ensemble mean. Further, similar to EMOS models for other weather or hydrological quantities, the exchangeable version of model (4.2.4) equals

$$\mu = a_0 + a_1\overline{f}_1 + \cdots + a_K\overline{f}_K, \qquad \sigma^2 = b_0 + b_1\overline{f}. \tag{4.2.5}$$

Note, that the expression of the mean (or location) as an affine function of the ensemble is general in EMOS post-processing (see e.g. Thorarinsdottir and Gneiting, 2010; Scheuerer, 2014; Baran and Lerch, 2015), whereas the dependence of the variance parameter on the ensemble mean is similar to the expression of the variance in the gamma BMA model of Sloughter *et al.* (2007) (see Section 4.1), and it is in line with the relation of forecast uncertainty to the expected precipitation amount mentioned in the introduction of this chapter. Moreover, practical tests show that, at least for the UWME and ALADIN-HUNEPS ensemble considered in the case studies of Section 4.3, models (4.2.4) and (4.2.5), respectively, significantly outperform the corresponding CSG EMOS models with variance parameters

$$\sigma^2 = b_0 + b_1 S^2 \qquad \text{and} \qquad \sigma^2 = b_0 + b_1\,\mathrm{MD},$$

where S^2 is the ensemble variance and

$$\mathrm{MD} := \frac{1}{K^2}\sum_{k,\ell=1}^{K} \big|f_k - f_\ell\big| \tag{4.2.6}$$

is the more robust ensemble mean difference (Scheuerer, 2014). Further, compared with the proposed models, natural modifications

$$\sigma^2 = b_0 + b_1 S^2 + b_2\overline{f} \qquad \text{or} \qquad \sigma^2 = (b_0 + b_1\overline{f})^2$$

in the CSG EMOS variance structure do not result in improved forecasts skills.

4.2.2 Censored generalized extreme value EMOS model

The EMOS model for precipitation accumulation proposed by Scheuerer (2014) is based on a censored GEV distribution $\mathcal{GEV}_0\big(\mu,\sigma,\xi\big)$ with location μ, scale $\sigma > 0$ and shape ξ having CDF

$$H_0(x|\,\mu,\sigma,\xi) = H(x|\,\mu,\sigma,\xi), \quad \text{if } x \geq 0, \quad \text{and} \quad H_0(x|\,\mu,\sigma,\xi) := 0, \quad \text{otherwise,} \tag{4.2.7}$$

where $H(x|\,\mu,\sigma,\xi)$ is defined by (3.2.10). For $-0.278 < \xi < 1$ this distribution has a positive skewness and an existing mean of

$$m = \begin{cases} \mu + \sigma\frac{\Gamma(1-\xi)-1}{\xi}, & \xi \neq 0; \\ \mu + \sigma\gamma, & \xi = 0, \end{cases}$$

where γ denotes the Euler-Mascheroni constant (see Section 3.2.6).

Scheuerer (2014) suggests to link the ensemble members to the mean and scale of the censored CSG distribution via

$$m = \alpha_0 + \alpha_1 f_1 + \cdots + \alpha_K f_K + \nu p_0 \qquad \text{and} \qquad \sigma = \beta_0 + \beta_1\,\mathrm{MD}, \tag{4.2.8}$$

where

$$p_0 := \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}_{\{f_k=0\}}$$

and MD is the ensemble mean difference defined by (4.2.6). In the exchangeable version of model (4.2.8) the link function to the mean is obviously

$$m = \alpha_0 + \alpha_1\overline{f}_1 + \cdots + \alpha_K\overline{f}_K + \nu p_0. \tag{4.2.9}$$

4.2.3 Parameter estimation

Following again the optimal score estimation principle of Gneiting and Raftery (2007), mean parameters $a_0, a_1, \ldots, a_K \geq 0$, variance parameters $b_0, b_1 \geq 0$ and shift parameter $\delta > 0$ of the CSG model of Section 4.2.1 can be estimated from the training data by optimizing the mean CRPS over the training set, and the same applies for the mean parameters $\alpha_0, \alpha_1, \ldots, \alpha_K \geq 0$, $\nu \in \mathbb{R}$, scale parameters $\beta_0, \beta_1 \geq 0$ and shape parameter $\xi \in (-0.278, 1)$ of the censored GEV model of Section 4.2.2.

Scheuerer and Hamill (2015) provide a closed expression for the CRPS for a CSG distribution in the form

$$\begin{aligned}\mathrm{CRPS}\big(\Gamma_0(\kappa,\theta,\delta),x\big) = {}& (x+\delta)\Big[2G(x+\delta|\kappa,\theta)-1\Big] - \frac{\theta\kappa}{\pi}\mathcal{B}\big(1/2,\kappa+1/2\big)\Big[1-G(2\delta|\,2\kappa,\theta)\Big]\\ &+\theta\kappa\Big[1+2G(\delta|\,\kappa,\theta)G(\delta|\,\kappa+1,\theta)-G^2(\delta|\,\kappa,\theta)-2G(y+\delta|\,\kappa+1,\theta)\Big]\\ &-\delta G^2(\delta|\,\kappa,\theta),\end{aligned}$$

where $\mathcal{B}$ denotes the beta function. CRPS of the censored GEV also has a simple closed form, which for $\xi \neq 0$ equals

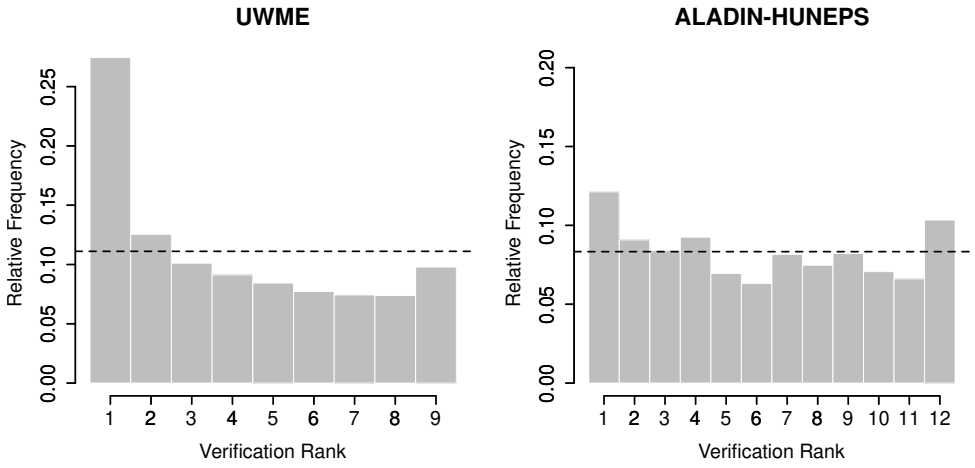
$$\begin{aligned}\mathrm{CRPS}\big(\mathcal{GEV}_0(\mu,\sigma,\xi),x\big) = {}& (\mu-x)\big[1-2H(x|\mu,\sigma,\xi)\big] + \mu H^2(0|\mu,\sigma,\xi)\\ &-2\frac{\sigma}{\xi}\Big[1-H(x|\mu,\sigma,\xi)-\Gamma_\ell\big(1-\xi,-\log H(x|\mu,\sigma,\xi)\big)\Big]\\ &+\frac{\sigma}{\xi}\Big[1-H^2(0|\mu,\sigma,\xi)-2^\xi\Gamma_\ell\big(1-\xi,-\log H(0|\mu,\sigma,\xi)\big)\Big],\end{aligned}$$

where Γ_ℓ stands for the lower incomplete gamma function, whereas for $\xi \in (-\varepsilon,\varepsilon)$ with a reasonably small ε, Scheuerer (2014) suggests to use approximation

$$\mathrm{CRPS}\big(\mathcal{GEV}_0(\mu,\sigma,\xi),x\big) \approx \frac{\varepsilon-\xi}{2\varepsilon}\mathrm{CRPS}\big(\mathcal{GEV}_0(\mu,\sigma,\varepsilon),x\big) + \frac{\varepsilon+\xi}{2\varepsilon}\mathrm{CRPS}\big(\mathcal{GEV}_0(\mu,\sigma,-\varepsilon),x\big).$$

4.3 Case studies

The predictive performance of the CSG EMOS model proposed by Baran and Nemoda (2016) (Section 4.2.1) is tested on ensemble forecasts produced by the UWME and ALADIN-HUNEPS EPSs, and the results are compared with the fits of the censored GEV EMOS (Section 4.2.1) and gamma BMA (Section 4.1) models investigated by Scheuerer (2014) and Sloughter *et al.* (2007), respectively, and the verification scores of the raw ensemble. We remark that according to the suggestions of Scheuerer (2014), for estimating



(a) (b)

Figure 4.1: Verification rank histograms (a) of the UWME forecasts for calendar year 2008 and (b) of the ALADIN-HUNEPS ensemble for the period 1 October 2010 – 25 March 2011.

the parameters of the GEV EMOS model for a given day, the estimates for the preceding day serve as initial conditions for the box constrained Broyden-Fletcher-Goldfarb-Shanno (Byrd *et al.*, 1995) optimization algorithm. Compared with the case of fixed initial conditions, this approach results in a slight increase of the forecast skills of the GEV EMOS model, whereas for the CSG EMOS method, at least in our case studies, fixed initial conditions are preferred. Further, for all investigated models we consider the regional approach to parameter estimation (see Section 1.3.3).

4.3.1 Data

University of Washington mesoscale ensemble

We consider 48 h forecasts of the 8-member UWME introduced in Section 3.3.1 and corresponding validating observations of 24 h precipitation accumulation for 152 stations in the Automated Surface Observing Network in five US states. The forecasts are initialized at 0 UTC, and similar to the case studies of Section 3.3 we investigate data for calendar year 2008 with additional forecasts and observations from the last three months of 2007 used for parameter estimation. After removing days and locations with missing data 83 stations remain resulting in 20 522 forecast cases for 2008.

Figure 4.1a shows the verification rank histogram of the raw ensemble, where zero observations are randomized among all zero forecasts. This histogram is far from the desired uniform distribution as in many cases the ensemble members overestimate the validating observation. The ensemble range contains the observed precipitation accumulation in 67.82 % of the cases, whereas the nominal coverage of the ensemble equals 77.78 %. Hence, the UWME is uncalibrated, and would require statistical post-processing to yield an improved forecast probability density function.

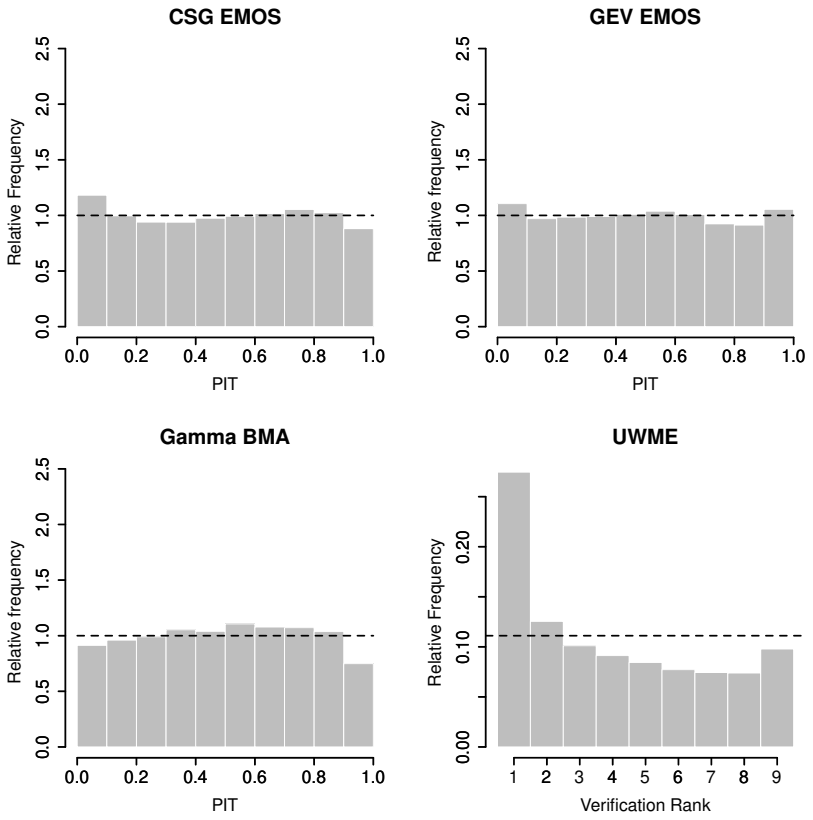


Figure 4.2: PIT histograms of EMOS and BMA post-processed UWME precipitation forecasts and verification rank histogram of the raw ensemble for calendar year 2008.

Model	CSG EMOS	GEV EMOS	Gamma BMA
Mean p-value	9.46×10^{-3}	2.49×10^{-2}	3.33×10^{-4}

Table 4.1: p-values of α^0_{1234} tests for uniformity of PIT values for the UWME. Means of 10000 random samples of sizes 2500 each.

ALADIN-HUNEPS ensemble

The ALADIN-HUNEPS precipitation data base at hand contains 11-member ensembles of 42 h forecasts (initialized at 18 UTC) produced by the ALADIN-HUNEPS system of the HMS (see Section 3.3.1) of 24 h precipitation accumulation for 10 major cities in Hungary (Budapest, Debrecen, Győr, Miskolc, Nagykanizsa, Nyíregyháza, Pécs, Sopron, Szeged, Szombathely) together with the corresponding validating observations for the period between 1 October 2010 and 25 March 2011. The data set is fairly complete since there are only two dates when three ensemble members are missing for all sites. These dates are excluded from the analysis.

The verification rank histogram of the raw ensemble, displayed in Figure 4.1b, shows far better calibration, than that of the UWME. The coverage of the ALADIN-HUNEPS ensemble equals 84.20 %, which is very close to the nominal value of 83.33 %.

Forecast	CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
CSG EMOS	2.252	3.019	80.46	8.350
GEV EMOS	2.283	3.033	79.91	8.683
Gamma BMA	2.357	3.220	83.44	9.515
Ensemble	2.929	3.708	67.95	8.599

Table 4.2: Mean CRPS of probabilistic forecasts, MAE of median forecasts and coverage and average width of 77.78 % central prediction intervals for the UWME.

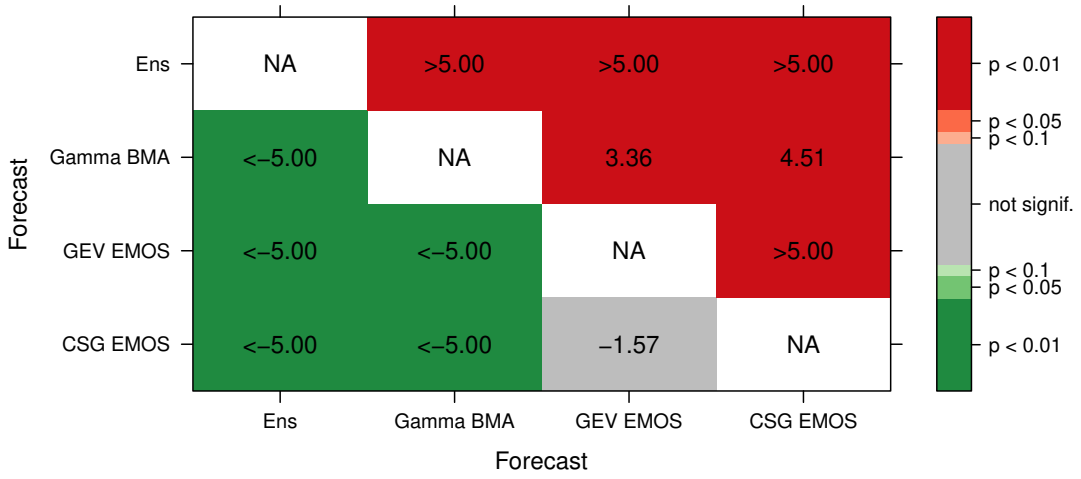


Figure 4.3: Values of the test statistic of the two-tailed DM test for equal predictive performance based on CRPS (*upper triangle*) and absolute error of median forecasts (*lower triangle*) for the UWME data. Green/red entries indicate superior performance of the forecast in the corresponding row/column.

4.3.2 Verification results for the UWME

The eight members of the UWME are generated using initial and boundary conditions from different sources, implying that the ensemble members are clearly distinguishable. Hence, similar to the corresponding wind speed models of Section 3.3, the mean and the variance of the underlying gamma distribution of the CSG EMOS model are linked to the ensemble members according to (4.2.4) with $K = 8$. Obviously, the reference censored GEV EMOS and gamma BMA models are also formulated under the assumption of non-exchangeable ensemble members.

A detailed study of CRPS and MAE values of the CSG EMOS and gamma BMA models corresponding to training period lengths of 20, 25, ..., 100 days indicates that both scores have global minima at 70 days. Hence, in our analysis we calibrate the UWME forecasts for calendar year 2008 using this training period length.

Figure 4.2 showing the PIT histograms of the CSG EMOS, GEV EMOS and gamma BMA models and the verification rank histogram of the raw ensemble clearly illustrates the advantage of statistical post-processing. Note that for our discrete-continuous models

Forecast	CRPSS	Brier Skill Score				
		0 mm	5 mm	15 mm	25 mm	30 mm
CSG EMOS	0.231	0.393	0.243	0.268	0.248	0.237
GEV EMOS	0.221	0.403	0.219	0.252	0.239	0.235
Gamma BMA	0.196	0.419	0.231	0.240	0.196	0.188

Table 4.3: CRPSS and BSS values with respect to the raw UWME.

in the case of zero observed precipitation the PIT is a random value chosen uniformly from the interval between zero and the probability of no precipitation (Sloughter *et al.*, 2007). Unfortunately, Knüppel's α^0_{1234} test rejects the uniformity of the PIT values for all models. However, for a quantification of the deviation from uniformity one can again consider the sampling approach of Section 3.3. Indeed, the mean p-values of 10000 random samples of PITs of sizes 2500 each, given in Table 4.1, nicely follow the shapes of the histograms of Figure 4.2. Note that the use of the Kolmogorov-Smirnov test results in the same ranking of the competing calibration methods. The mean p-values for CSG EMOS, GEV EMOS and Gamma BMA models are 0.154, 0.310 and 0.044, respectively.

In Table 4.2 the mean CRPS of probabilistic forecasts, the MAE of median forecasts and the coverage and average width of 77.78 % central prediction intervals for the two EMOS approaches, the gamma BMA model and the raw ensemble are reported, whereas Figure 4.3 shows the results of DM tests for equal predictive performance based on the CRPS values and the absolute errors of median forecasts. By examining these results, one can clearly observe the obvious advantage of post-processing with respect to the raw ensemble, which is quantified in the significant decrease of the mean CRPS and MAE values and in a substantial improvement in coverage. Further, the CSG EMOS model results in the lowest mean CRPS, whereas in terms MAE there is no difference between the two EMOS methods, which significantly outperform the gamma BMA approach both in calibration of probabilistic and accuracy of point forecasts. The CSG EMOS model results in the sharpest central prediction interval combined with a rather fair coverage, whereas the central prediction intervals corresponding to the other two calibration methods are slightly wider than that of the raw ensemble.

The improvement in calibration caused by statistical post-processing can also be observed in skill scores reported in Table 4.3 and reliability diagrams displayed in Figure 4.4. Note that thresholds 5, 15, 25, 30 mm correspond approximately to the 45th, 75th, 85th and 90th percentiles of the observed non-zero precipitation accumulation. Gamma BMA method performs well in predicting the probability of positive precipitation and exceeding the 5 mm threshold, whereas for higher threshold values it is behind the two EMOS approaches, where the CSG EMOS model presents slightly better forecast skills. Hence, one can conclude, that in case of the UWME the EMOS approaches outperform both the raw ensemble and the gamma BMA model and the proposed CSG EMOS model slightly outperforms the GEV EMOS method.

4.3.3 Verification results for the ALADIN-HUNEPS ensemble

As a contrast to the UWME, the way the ALADIN-HUNEPS ensemble is generated induces a natural grouping of the ensemble members. The first group contains the control,

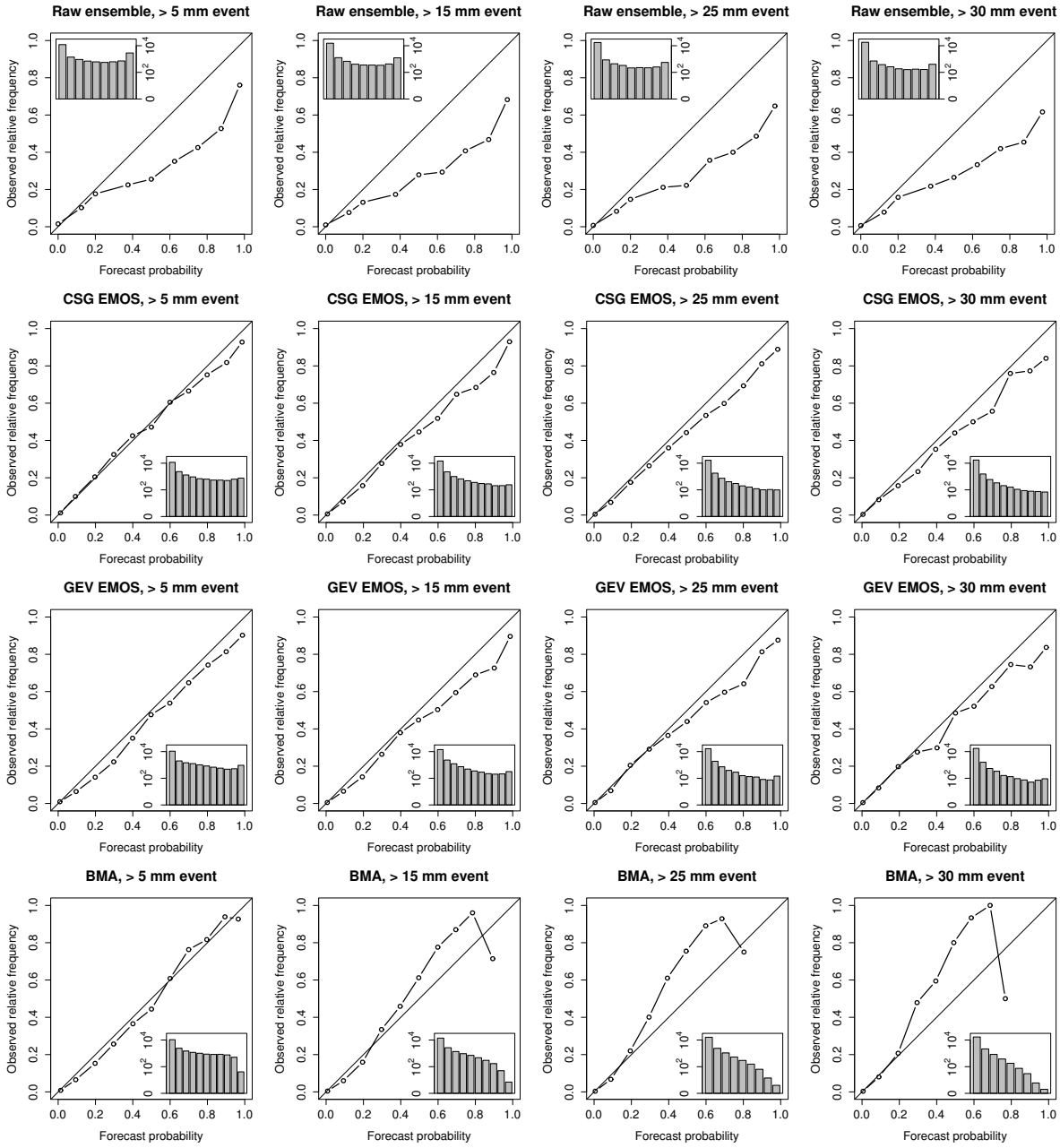


Figure 4.4: Reliability diagrams of the raw ensemble and EMOS and BMA post-processed forecasts for the UWME for the calendar year 2008. The inset histograms display the log-frequency of cases within the respective bins.

whereas the second group consists of the 10 exchangeable ensemble members. This splitting results in the GEV EMOS model (4.2.5) with $K = 2$, $M_1 = 1$ and $M_2 = 10$, and the same grouping is considered for the benchmark GEV EMOS and gamma BMA models.

Again, in order to determine the appropriate length of the rolling training period the mean CRPS and MAE values of the various models for training periods of lengths $20, 25, \ldots, 100$ calendar days are investigated. In order to ensure the comparability of the results corresponding to different training period lengths, verification scores from 10

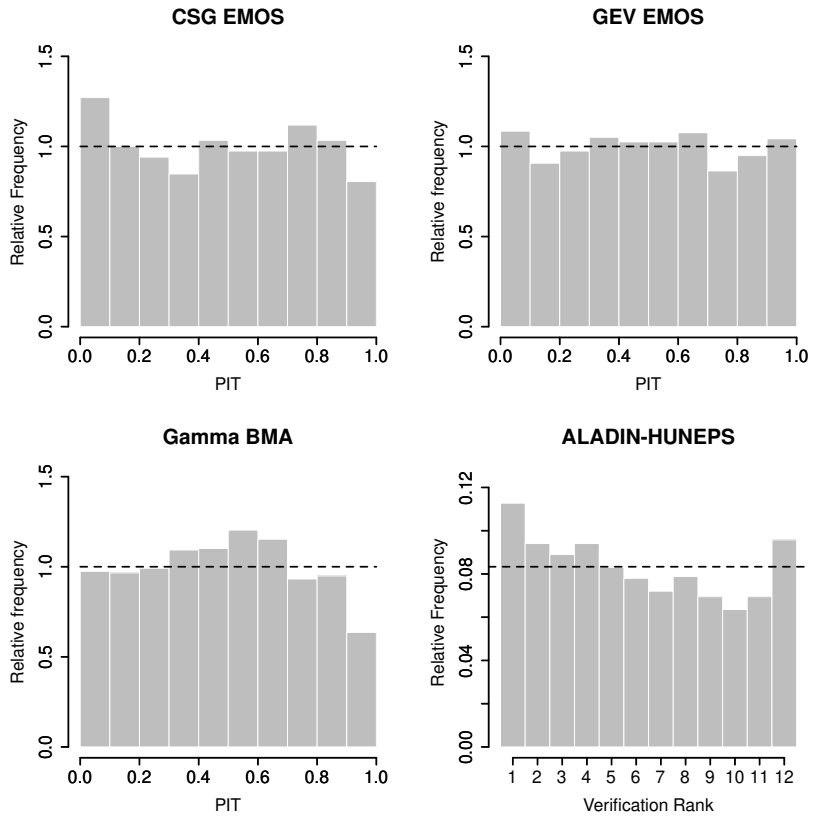


Figure 4.5: PIT histograms of EMOS and BMA post-processed ALADIN-HUNEPS ensemble forecasts of precipitation and verification rank histogram of the raw ensemble for the period 27 November 2010 – 25 March 2011.

Model	CSG EMOS	GEV EMOS	Gamma BMA
p-value	2.90×10^{-3}	0.907	2.21×10^{-5}

Table 4.4: p-values of α^0_{1234} tests for uniformity of PIT values for the ALADIN-HUNEPS ensemble.

January to 25 March 2011 are considered. The corresponding curves of the CRPS and MAE scores plotted against the training period lengths (not shown) have global minima at 85 days, however they have elbows at 55 days, that is, up to this training period length the decrease is rather steep then the values stabilize. Hence, as in general shorter training periods are preferred, for calibrating the ALADIN-HUNEPS ensemble a training period of length 55 days is used. This means that ensemble members, validating observations, and predictive PDFs are available for the period from 27 November 2010 to 25 March 2011 having 119 calendar days (just after the first 55 day training period) and 1 180 forecast cases, since on 15 February 2011 three ensemble members are missing and this date is excluded from the analysis. This time interval starts more than 6 weeks earlier than the one used for determination of the optimal training period length.

Compared with the verification rank histogram of the raw ensemble, the PIT histograms of the post-processed forecasts displayed in Figure 4.5 show a substantial improvement in calibration. For the GEV EMOS model the α^0_{1234} test accepts the uniformity

Forecast	CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
CSG EMOS	0.465	0.636	89.15	2.185
GEV EMOS	0.477	0.641	86.53	2.192
Gamma BMA	0.532	0.708	93.73	2.854
Ensemble	0.485	0.640	84.24	2.436

Table 4.5: Mean CRPS of probabilistic forecasts, MAE of median forecasts and coverage and average width of 83.33 % central prediction intervals for the ALADIN-HUNEPS ensemble.

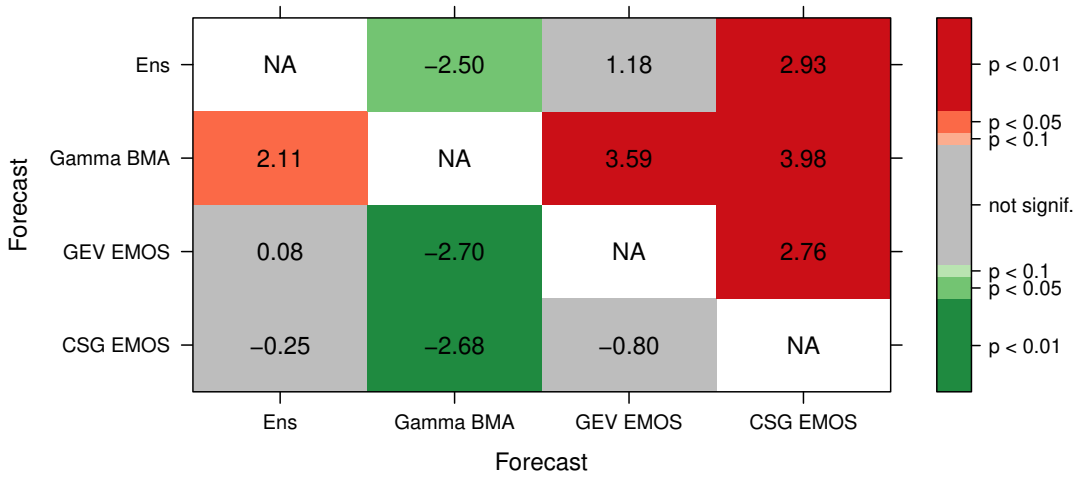


Figure 4.6: Values of the test statistic of the two-tailed DM test for equal predictive performance based on CRPS (*upper triangle*) and absolute error of median forecasts (*lower triangle*) for the ALADIN-HUNEPS data. Green/red entries indicate superior performance of the forecast in the corresponding row/column.

of the PIT values (see Table 4.4 and note the extremely high p-value), whereas the other two p-values are in accordance with the slight bias of the histogram of the CSG EMOS and the hump shaped histogram of the Gamma BMA model indicating some overdispersion. Note that based on the Kolmogorov-Smirnov test, PITs of both EMOS models can be taken as uniformly distributed. The p-values for the CSG and GEV EMOS and the gamma BMA are 0.119, 0.921 and 0.003, respectively.

Concerning the two EMOS approaches, the verification scores of Table 4.5 together with the results of the corresponding DM tests for equal predictive performance (see Figure 4.6) display similar behaviour as in the case of the UWME. There is no significant difference between the MAE values of the CSG and GEV EMOS methods and the former results in the lowest CRPS and the sharpest 83.33 % central prediction interval. Further, the EMOS models significantly outperform both the raw ensemble and the gamma BMA approach, despite the raw ensemble is rather well calibrated and has far better predictive skill than the BMA calibrated forecast. Note that the large mean CRPS and coverage

Forecast	CRPSS	Brier Skill Score				
		0 mm	1 mm	5 mm	7 mm	9 mm
CSG EMOS	0.042	0.094	0.057	-0.011	-0.025	0.019
GEV EMOS	0.017	0.166	0.008	-0.022	-0.030	0.027
Gamma BMA	-0.098	0.151	-0.070	-0.265	-0.136	-0.023

Table 4.6: CRPSS and BSS values with respect to the raw ALADIN-HUNEPS ensemble.

of the BMA predictive distribution is totally in line with the shape of the corresponding PIT histogram of Figure 4.5.

The good predictive performance of the ALADIN-HUNEPS ensemble can also be observed on the large amount of negative skill scores reported in Table 4.6 (threshold values 1, 5, 7, 9 mm again correspond approximately to the 45th, 75th, 85th and 90th percentiles of the observed non-zero precipitation accumulation) and on the reliability diagrams of Figure 4.7. Similar to the case of the UWME, for 0 mm threshold the gamma BMA model has good predictive performance, whereas for higher threshold values it underperforms the CSG and GEV EMOS models and the raw ensemble. However, in connection with the reliability diagrams one should also note that the hectic behaviour of the graphs (compared with the rather smooth diagrams of Figure 4.4) is a consequence of the shortage of data, as the verification period contains only 394 observations of positive precipitation, which is around one third of the forecast cases.

Taking into account both the uniformity of the PIT values and the verification scores in Tables 4.5 and 4.6 it can be said that the proposed CSG EMOS model has the best overall performance in calibration of the raw ALADIN-HUNEPS ensemble forecasts of precipitation accumulation.

4.3.4 Computational aspects

As it has been already mentioned in the case studies of Section 3.3, in EMOS modelling the numerical optimization used in parameter estimation, whereas in BMA calibration the EM algorithm including also an optimization step is the most time consuming part. In Figures 4.8a and 4.8b the kernel density estimates of the distribution of computation times over the days in the verification period for the competing post-processing models are plotted for UWME and ALADIN-HUNEPS ensemble, respectively. Modelling was performed on the same portable computer as in the other case studies (Intel Quad Core i7-4700MQ CPU (2.40GHz $\times$ 4), 20 Gb RAM) with the help of the `ensembleMOS` (Yuen *et al.*, 2018) and `ensembleBMA` (Fraley *et al.*, 2011) packages of `R`. The GEV EMOS approach outperforms the CSG EMOS in terms of computation costs and both EMOS methods provide faster modelling than the BMA. However, even the longest estimation procedure calculating BMA model parameters for UWME for a given day took 148 seconds, so all three investigated models seem to be fast enough for operational use.

4.4 Conclusions

In this chapter we describe a new EMOS model for calibrating ensemble forecasts of precipitation accumulation, where the predictive distribution follows a censored and shifted

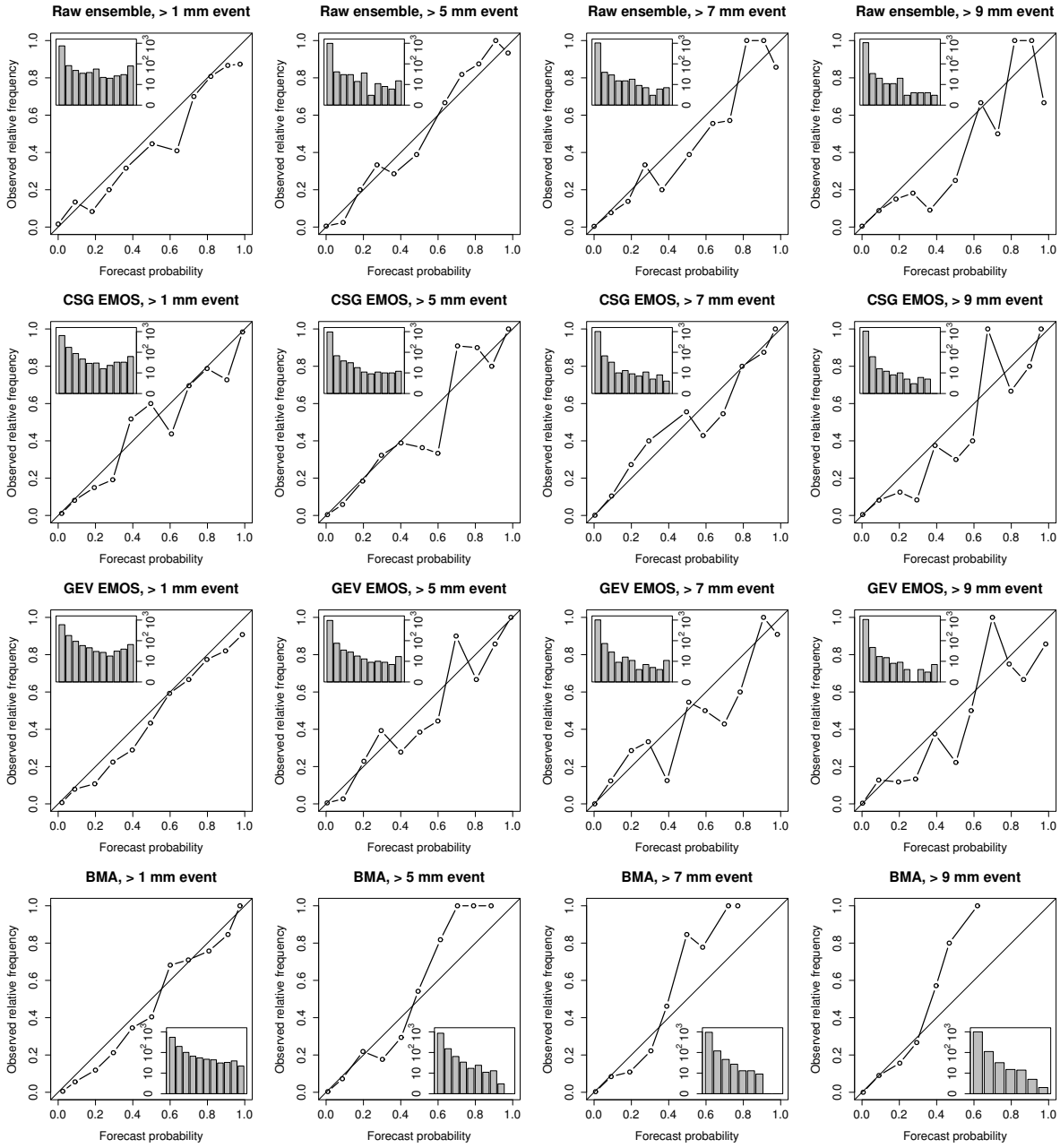
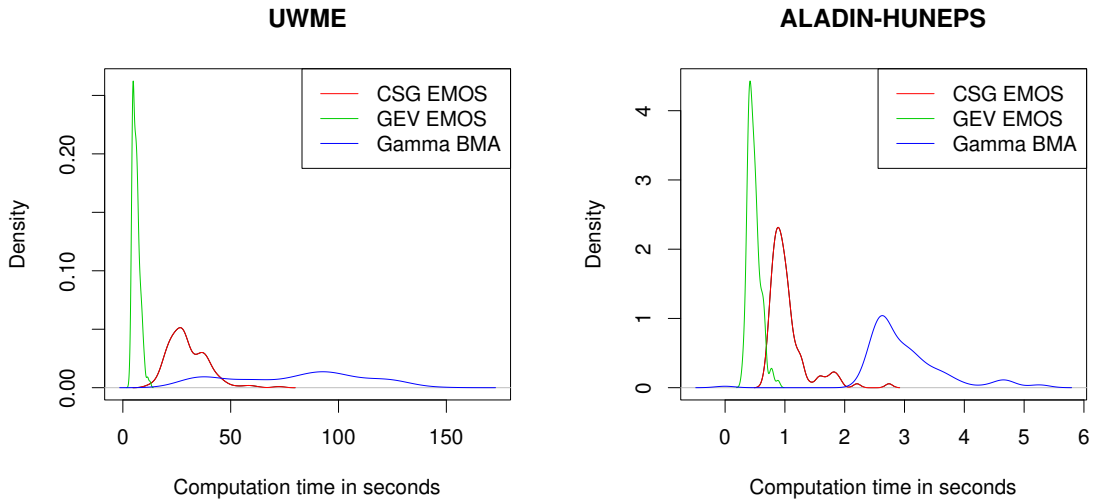


Figure 4.7: Reliability diagrams of the raw ensemble and EMOS and BMA post-processed forecasts for the ALADIN-HUNEPS ensemble for the period 27 November 2010 – 25 March 2011. The inset histograms display the log-frequency of cases within the respective bins.

gamma distribution, with mean and variance of the underlying gamma law being affine functions of the raw ensemble and the ensemble mean, respectively. The CSG EMOS method is tested on ensemble forecasts of 24 h precipitation accumulation of the 8-member University of Washington mesoscale ensemble and on the 11-member ALADIN-HUNEPS ensemble of the Hungarian Meteorological Service. These ensemble prediction systems differ both in the climate of the covered area and in the generation of the ensemble members. By investigating the uniformity of the PIT values of predictive distributions, the mean CRPS of probabilistic forecasts, the Brier scores and reliability diagrams for various



(a) (b)

Figure 4.8: Densities of computation times for the CSG EMOS, GEV EMOS and gamma BMA models. (a) UWME for the calendar year 2008; (b) ALADIN-HUNEPS ensemble for the period 27 November 2010 – 25 March 2011.

thresholds, the MAE of median forecasts and the average width and coverage of central prediction intervals corresponding to the nominal coverage, the predictive skill of the new approach is compared with that of the GEV EMOS method (Scheuerer, 2014), the gamma BMA model (Sloughter *et al.*, 2007) and the raw ensemble. From the results of the presented case studies one can conclude that in terms of calibration of probabilistic and accuracy of point forecasts the proposed CSG EMOS model significantly outperforms both the raw ensemble and the BMA model and shows slightly better forecast skill than the GEV EMOS approach.

Chapter 5

Bivariate models for wind speed and temperature

As it has already been mentioned in Section 1.3, for temperature observations normal BMA and EMOS models fit reasonably well, while for wind speed observations BMA methods with gamma and truncated normal components and EMOS approaches based on truncated normal, log-normal and generalized extreme value distributions have been developed. This gives the natural idea of joint modelling wind speed and temperature with a bivariate normal distribution with first (wind) coordinate truncated from below at zero. Based on Baran and Möller (2015) and Baran and Möller (2017), respectively, in this chapter we introduce a bivariate BMA and a bivariate EMOS model for joint calibration of ensemble forecasts of these two weather quantities. The predictive performance of the new bivariate approaches is tested on two data sets based on UWME and ALADIN-HUNEPS ensemble prediction systems.

5.1 Bivariate BMA model

The proposed BMA approach (Baran and Möller, 2015) is based on a bivariate truncated normal distribution with first coordinate truncated from below at zero $\mathcal{N}_2^0(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_W \\ \mu_T \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \sigma_W^2 & \sigma_{WT} \\ \sigma_{WT} & \sigma_T^2 \end{bmatrix}$$

denote the location vector and scale matrix, respectively. Along this chapter subscripts W and T refer to wind speed and temperature, respectively. If Σ is regular, the joint PDF of this special bivariate distribution equals

$$g(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) := \frac{\big(\det(\Sigma)\big)^{-1/2}}{2\pi\Phi\big(\mu_W/\sigma_W\big)} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big)\mathbb{I}_{\{x_W \geq 0\}}, \quad \boldsymbol{x} = \begin{bmatrix} x_W \\ x_T \end{bmatrix} \in \mathbb{R}^2, \tag{5.1.1}$$

whereas the corresponding mean vector $\boldsymbol{\kappa}$ and covariance matrix Ξ are

$$\boldsymbol{\kappa} = \boldsymbol{\mu} + \frac{\varphi\big(\mu_W/\sigma_W\big)}{\Phi\big(\mu_W/\sigma_W\big)} \begin{bmatrix} \sigma_W \\ \sigma_{WT}/\sigma_W \end{bmatrix} \qquad \text{and}$$
$$\Xi = \Sigma - \left(\frac{\mu_W}{\sigma_W} \frac{\varphi\big(\mu_W/\sigma_W\big)}{\Phi\big(\mu_W/\sigma_W\big)} + \left(\frac{\varphi\big(\mu_W/\sigma_W\big)}{\Phi\big(\mu_W/\sigma_W\big)} \right)^2 \right) \begin{bmatrix} \sigma_W^2 & \sigma_{WT} \\ \sigma_{WT} & \sigma_{WT}^2/\sigma_W^2 \end{bmatrix},$$

respectively.

5.1.1 Model formulation

Consider a BMA mixture (1.3.1) where the location vector $\boldsymbol{\mu}_k$ of the kth component PDF is an affine function of the corresponding ensemble member $\boldsymbol{f}_k$ and the scale matrices of all components are equal, resulting in the model

$$p(\boldsymbol{x}|\, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_K; A_1, \ldots, A_K; B_1, \ldots, B_K; \Sigma) := \sum_{k=1}^{K} \omega_k g(\boldsymbol{x}|\, A_k + B_k \boldsymbol{f}_k, \Sigma), \tag{5.1.2}$$

where g is the PDF defined by (5.1.1), $A_k \in \mathbb{R}^2$ and B_k is a two-by-two real matrix. In this way model (5.1.2) is a direct extension of the univariate BMA models of temperature and wind speed investigated in Raftery *et al.* (2005) and Baran (2014), where the authors also used the assumption of a common scale parameter for all BMA components. It reduces the number of parameters to be estimated and makes computations easier.

One can have an even more parsimonious model by using the same bias correction parameters for all ensemble members, resulting in the predictive PDF

$$q(\boldsymbol{x}|\, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_K; A; B; \Sigma) := \sum_{k=1}^{K} \omega_k g(\boldsymbol{x}|\, A + B \boldsymbol{f}_k, \Sigma). \tag{5.1.3}$$

We remark that a similar type of simplification is used in the wind speed model of the `ensembleBMA` package of `R` (Fraley *et al.*, 2011).

5.1.2 Parameter estimation

Similar to the univariate BMA approaches, model parameters A_k, B_k, ω_k, $k = 1, 2, \ldots, K$, and Σ of PDF (5.1.2) and A, B, Σ and ω_k, $k = 1, 2, \ldots, K$, of PDF (5.1.3) are usually estimated using training data consisting of ensemble members and validating observations from the preceding n days (rolling training period). In what follows, $\boldsymbol{f}_{k,s,t}$ denotes the kth ensemble member vector for location $s \in \mathcal{S}$ and time $t \in \mathcal{T}$, and by $\boldsymbol{x}_{s,t}$ we denote the corresponding validating observation. Here we consider a bivariate generalization of the pure ML approach with EM algorithm for truncated normal mixtures described in Section 2.1.2, see also Baran *et al.* (2019a) and Baran (2014).

Full model

Under the assumption of independence of forecast errors in space and time, the log-likelihood function corresponding to model (5.1.2) equals

$$\ell(\omega_1,\ldots,\omega_M;A_1,\ldots,A_K;B_1,\ldots,B_K;\Sigma)=\sum_{s,t}\log\left[\sum_{k=1}^{K}\omega_k g\big(\boldsymbol{x}_{s,t}|\,A_k+B_k\boldsymbol{f}_{k,s,t},\Sigma\big)\right], \tag{5.1.4}$$

where the first summation is over all locations $s\in\mathcal{S}$ and time points t from the training period containing N forecast cases (N distinct values of (s,t)).

After introducing latent allocation variables $z_{k,s,t}$ taking values one or zero according as whether $\boldsymbol{x}_{s,t}$ comes from the kth component PDF or not, the complete data log-likelihood corresponding to the training data and allocations equals

$$\ell_C(\omega_1,\ldots,\omega_K;A_1,\ldots,A_K;B_1,\ldots,B_K;\Sigma) = \sum_{s,t}\sum_{k=1}^{K} z_{k,s,t}\Big[\log(\omega_k)+\log\Big(g\big(\boldsymbol{x}_{s,t}|\,A_k+B_k\boldsymbol{f}_{k,s,t},\Sigma\big)\Big)\Big].$$

As mentioned in Section 2.1.2, the EM algorithm starts with initial values of the parameters, then alternates between an expectation (E) step and a maximization (M) step until convergence. The coefficients of linear regression of the validating observations on the corresponding ensemble members can serve as initial values of $A_k^{(0)}$ and $B_k^{(0)}$, $k=1,2,\ldots,K$, the covariance matrix of the validating observations can be taken as $\Sigma^{(0)}$, while the initial weights $\omega_k^{(0)}$, $k=1,2,\ldots,K$, might be set to be all equal.

For the truncated normal mixture model given by (5.1.1) and (5.1.2) the E step is,

$$z_{k,s,t}^{(j+1)}:=\frac{\omega_k^{(j)}g\big(\boldsymbol{x}_{s,t}|\,A_k^{(j)}+B_k^{(j)}\boldsymbol{f}_{k,s,t},\Sigma^{(j)}\big)}{\sum_{i=1}^{M}\omega_i^{(j)}g\big(\boldsymbol{x}_{s,t}|\,A_i^{(j)}+B_i^{(j)}\boldsymbol{f}_{i,s,t},\Sigma^{(j)}\big)}, \tag{5.1.5}$$

where the superscript refers to the actual iteration. Observe again, that the above estimates of $z_{k,s,t}$ are usually not integers even though the true values of these latent allocation variables are either 0 or 1. Further, the first part of the M step is

$$\omega_k^{(j+1)}:=\frac{1}{N}\sum_{s,t}z_{k,s,t}^{(j+1)}, \tag{5.1.6}$$

while the second part can be derived from equations

$$\frac{\partial\ell_C}{\partial A_k}=0,\qquad \frac{\partial\ell_C}{\partial B_k}=0,\qquad \frac{\partial\ell_C}{\partial\Sigma}=0,\qquad k=1,2,\ldots,K. \tag{5.1.7}$$

As the above system of equations is non-linear, we suggest iteration steps

$$A_k^{(j+1)} := \left[\sum_{s,t} z_{k,s,t}^{(j+1)} \left(\left(\boldsymbol{x}_{s,t} - B_k^{(j)} \boldsymbol{f}_{k,s,t}\right) - \frac{1}{\sigma_W^{(j)}} \frac{\varphi\left(\mu_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)}{\Phi\left(\mu_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)} \begin{bmatrix} \left(\sigma_W^{(j)}\right)^2 \\ \sigma_{WT}^{(j)} \end{bmatrix} \right) \right] \left[\sum_{s,t} z_{k,s,t}^{(j+1)}\right]^{-1},$$

$$B_k^{(j+1)} := \left[\sum_{s,t} z_{k,s,t}^{(j+1)} \left(\left(\boldsymbol{x}_{s,t} - A_k^{(j+1)}\right) - \frac{1}{\sigma_W^{(j)}} \frac{\varphi\left(\tilde{\mu}_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)}{\Phi\left(\tilde{\mu}_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)} \begin{bmatrix} \left(\sigma_W^{(j)}\right)^2 \\ \sigma_{WT}^{(j)} \end{bmatrix} \right) \boldsymbol{f}_{k,s,t}^\top \right] \tag{5.1.8}$$

$$\times \left[\sum_{s,t} z_{k,s,t}^{(j+1)} \boldsymbol{f}_{k,s,t} \boldsymbol{f}_{k,s,t}^\top\right]^{-1},$$

$$\Sigma^{(j+1)} := \frac{1}{N} \sum_{s,t} \sum_{k=1}^{K} z_{k,s,t}^{(j+1)} \left(\left(\boldsymbol{x}_{s,t} - \boldsymbol{\mu}_{k,s,t}^{(j+1)}\right)\left(\boldsymbol{x}_{s,t} - \boldsymbol{\mu}_{k,s,t}^{(j+1)}\right)^\top \right.$$

$$\left. + \boldsymbol{\mu}_{k,s,t}^{(j+1)} \frac{1}{\sigma_W^{(j)}} \frac{\varphi\left(\mu_{W,k,s,t}^{(j+1)}/\sigma_W^{(j)}\right)}{\Phi\left(\mu_{W,k,s,t}^{(j+1)}/\sigma_W^{(j)}\right)} \begin{bmatrix} \left(\sigma_W^{(j)}\right)^2 & \sigma_{WT}^{(j)} \\ \sigma_{WT}^{(j)} & \left(\sigma_{WT}^{(j)}/\sigma_W^{(j)}\right)^3 \end{bmatrix} \right),$$

where $\mu_{W,k,s,t}^{(j)}$ and $\tilde{\mu}_{W,k,s,t}^{(j)}$ denote the first (wind) coordinates of $\boldsymbol{\mu}_{k,s,t}^{(j)} := A_k^{(j)} + B_k^{(j)} \boldsymbol{f}_{k,s,t}$ and $\tilde{\boldsymbol{\mu}}_{k,s,t}^{(j)} := A_k^{(j+1)} + B_k^{(j)} \boldsymbol{f}_{k,s,t}$, respectively.

Parsimonious model

For the parsimonious model (5.1.3) the log-likelihood function is obviously

$$\ell(\omega_1, \ldots, \omega_K; A; B; \Sigma) = \sum_{s,t} \log \left[\sum_{k=1}^{K} \omega_k g\left(\boldsymbol{x}_{s,t} \middle| A + B\boldsymbol{f}_{k,s,t}, \Sigma\right)\right],$$

which is maximized using the same type of EM algorithm as before. The E step, and the iterations corresponding to $\omega_k^{(j+1)}$ and $\Sigma^{(j+1)}$ are obvious modifications of (5.1.5), (5.1.6) and of the last iteration of (5.1.8), respectively, while the first two iterations of (5.1.8) should be replaced by

$$A^{(j+1)} := \frac{1}{N} \sum_{s,t} \sum_{k=1}^{K} z_{k,s,t}^{(j+1)} \left(\left(\boldsymbol{x}_{s,t} - B^{(j)} \boldsymbol{f}_{k,s,t}\right) - \frac{1}{\sigma_W^{(j)}} \frac{\varphi\left(\mu_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)}{\Phi\left(\mu_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)} \begin{bmatrix} \left(\sigma_W^{(j)}\right)^2 \\ \sigma_{WT}^{(j)} \end{bmatrix} \right),$$

$$B^{(j+1)} := \sum_{s,t} \sum_{k=1}^{K} z_{k,s,t}^{(j+1)} \left(\left(\boldsymbol{x}_{s,t} - A^{(j+1)}\right) - \frac{1}{\sigma_W^{(j)}} \frac{\varphi\left(\tilde{\mu}_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)}{\Phi\left(\tilde{\mu}_{W,k,s,t}^{(j)}/\sigma_W^{(j)}\right)} \begin{bmatrix} \left(\sigma_W^{(j)}\right)^2 \\ \sigma_{WT}^{(j)} \end{bmatrix} \right) \boldsymbol{f}_{k,s,t}^\top$$

$$\times \left[\sum_{s,t} \sum_{k=1}^{K} z_{k,s,t}^{(j+1)} \boldsymbol{f}_{k,s,t} \boldsymbol{f}_{k,s,t}^\top\right]^{-1}.$$

In this case $\mu_{W,k,s,t}^{(j)}$ and $\tilde{\mu}_{W,k,s,t}^{(j)}$ denote the first coordinates of $\boldsymbol{\mu}_{k,s,t}^{(j)} := A^{(j)} + B^{(j)} \boldsymbol{f}_{k,s,t}$ and $\tilde{\boldsymbol{\mu}}_{k,s,t}^{(j)} := A^{(j+1)} + B^{(j)} \boldsymbol{f}_{k,s,t}$, respectively.

5.2 Bivariate truncated normal EMOS model

5.2.1 Model formulation

As a simple alternative to joint BMA post-processing of wind speed and temperature forecasts, one can consider an EMOS approach (Baran and Möller, 2017) with bivariate predictive distribution

$$\mathcal{N}_2^0\big(\mathcal{A}+\mathcal{B}_1\boldsymbol{f}_1+\cdots+\mathcal{B}_K\boldsymbol{f}_K,\mathcal{C}+\mathcal{D}\boldsymbol{S}\mathcal{D}^\top\big) \quad \text{with} \quad \boldsymbol{S}:=\frac{1}{K-1}\sum_{k=1}^{K}\big(\boldsymbol{f}_k-\overline{\boldsymbol{f}}\big)\big(\boldsymbol{f}_k-\overline{\boldsymbol{f}}\big)^\top, \tag{5.2.1}$$

where $\overline{\boldsymbol{f}}$ denotes the ensemble mean vector. Parameter vector $\mathcal{A}\in\mathbb{R}^2$ and two-by-two real parameter matrices $\mathcal{B}_1,\ldots,\mathcal{B}_K$ and $\mathcal{C}$, $\mathcal{D}$ of model (5.2.1), where $\mathcal{C}$ is assumed to be symmetric and non-negative definite, can again be estimated e.g. from rolling training data. According to the general procedure for EMOS models, the estimates optimize the mean of a proper verification score over all forecast cases of the training set. Here we optimize the mean logarithmic score (1.4.3), and we remark again that under the assumption of independence in space and time this approach is equivalent to the ML method. Obviously, the forecast errors are usually not independent, however, since one is estimating the conditional distribution of a single weather quantity vector with respect to the corresponding forecasts, the parameter estimates are not really sensitive to this assumption (see e.g. Raftery *et al.*, 2005).

In the case of existence of groups of exchangeable ensemble members, one has to follow the usual procedure and instead of model (5.2.1) consider the predictive distribution

$$\mathcal{N}_2^0\big(\mathcal{A}+\mathcal{B}_1\overline{\boldsymbol{f}}_1+\cdots+\mathcal{B}_K\overline{\boldsymbol{f}}_K,\mathcal{C}+\mathcal{D}\boldsymbol{S}\mathcal{D}^\top\big), \tag{5.2.2}$$

where $\overline{\boldsymbol{f}}_k$ denotes the mean of the kth group.

5.2.2 Parameter estimation

In bivariate EMOS models (5.2.1) and (5.2.2) the number of free parameters to be estimated is $4K+10$, which means 14 unknown parameters even in the simplest case of a single exchangeable ensemble group. Hence, for estimating the parameters of models (5.2.1) and (5.2.2) mostly the regional EMOS approach (see Section 1.3.3) is applicable, unless one has an extremely large data set allowing very long training periods.

The mean logarithmic score is optimized numerically using principally the Nelder-Mead algorithm, as the faster but less robust BFGS becomes unstable in the case of a small training set. Both optimization methods require initial values, and the starting values of the location parameters $\mathcal{A}$ and $\mathcal{B}_1,\ldots,\mathcal{B}_K$ are coefficients of the bivariate linear regression of the observations on the ensemble forecasts over the training period. Further, for the scale parameters $\mathcal{C}$ and $\mathcal{D}$, the previous day's estimates can serve as initials values, however, according to our experience, fixed starting values (we simply use two-by-two unit matrices) provide slightly better results. Finally, to enforce the non-negative definiteness of the parameter matrix C, one can set $\mathcal{C}=CC^\top$ and perform the optimization with respect to C.

5.3 Gaussian copula approach

The Gaussian copula approach allows to construct a post-processed joint distribution based on the individually post-processed marginals. For d weather variables of interest with (individually post-processed) marginal distributions $F_1, F_2, \ldots, F_d$, the joint distribution $\boldsymbol{F}$ of the weather variables takes the following form under a Gaussian copula model

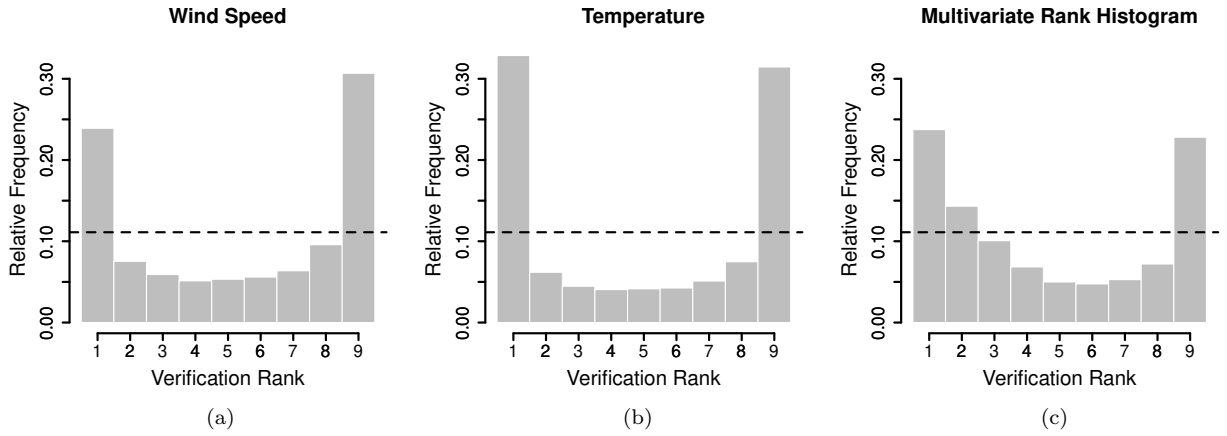
$$\boldsymbol{F}(x_1, \ldots, x_d \,|\, \boldsymbol{C}) := \Phi_d\big(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d)) \,\big|\, \boldsymbol{C}\big).$$

Here, Φ^{-1} denotes the inverse CDF of a standard Gaussian distribution, $\Phi_d(\cdot \,|\, \Sigma)$ is the CDF of a d-variate Gaussian distribution with covariance matrix Σ, whereas $\boldsymbol{C}$ is a $d \times d$ correlation matrix, i.e. a positive definite matrix with unit diagonal. To be fully defined, the Gaussian copula requires only the marginal distributions $F_1, F_2, \ldots, F_d$ and the correlation matrix $\boldsymbol{C}$. For univariate post-processing of the marginal distributions $F_1, F_2, \ldots, F_d$ any post-processing model of choice can be used. In the original approach of Möller *et al.* (2013) the marginals were post-processed merely with suitable BMA models, whereas for the comparison with the bivariate methods presented in Sections 5.1 and 5.2, the copula marginals are fitted with appropriate univariate BMA and EMOS models, respectively. While each observation is associated with its own copula $\boldsymbol{F}$, they all share the same correlation matrix. Therefore, $\boldsymbol{C}$ can be obtained by estimating latent Gaussian factors $z_j = \Phi^{-1}\big(F_j(x_j)\big)$, $j = 1, 2, \ldots, d$, from observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ of a separate (historic) data set. The correlation matrix is then directly estimated from the fitted latent Gaussian factors, for further details see Möller *et al.* (2013).

5.4 Case studies

The forecast skill of the bivariate BMA and EMOS models described in Sections 5.1.1 and 5.2, respectively, is tested on the 8-member UWME and on the ALADIN-HUNEPS ensemble of the HMS. Model parameters in both case studies are estimated using the global approach (see Section 1.3.3). The goodness of fit of the predictive distributions is quantified with the multivariate scores given in Section 1.4, and the obtained results are compared with the fits of the independent BMA and EMOS models of wind speed (Baran, 2014; Thorarinsdottir and Gneiting, 2010) and temperature (Raftery *et al.*, 2005; Gneiting *et al.*, 2005) and the Gaussian copula method proposed by Möller *et al.* (2013) both with BMA and EMOS marginal distributions. We remark that the parameters of the independent univariate EMOS models are estimated by minimizing the mean CRPS of the training data. For fitting the marginal predictive distributions in the Gaussian copula approach, we employ the same univariate BMA and EMOS models for wind speed and temperature as in the independent case. Therefore, their model parameters are estimated by the minimum CRPS method as well. If one has a closed expression for the CRPS, which is the case both for the normal and the truncated normal distribution, this method usually gives better results than optimization with respect to the logarithmic score.

Further, for the case study conducted in Möller *et al.* (2013), the univariate post-processing of the copula marginals is performed at each considered station individually, as the performance of the method at specific stations as well as the structure of correlations were investigated. Here the copula marginals are formed by applying global BMA and EMOS models to have a better comparability to the proposed bivariate approaches. This



(a) (b) (c)

Figure 5.1: Verification rank histograms of the UMWE forecasts of maximum wind speed (a) and minimum temperature (b) and the multivariate rank histogram (c) for the calendar year 2008.

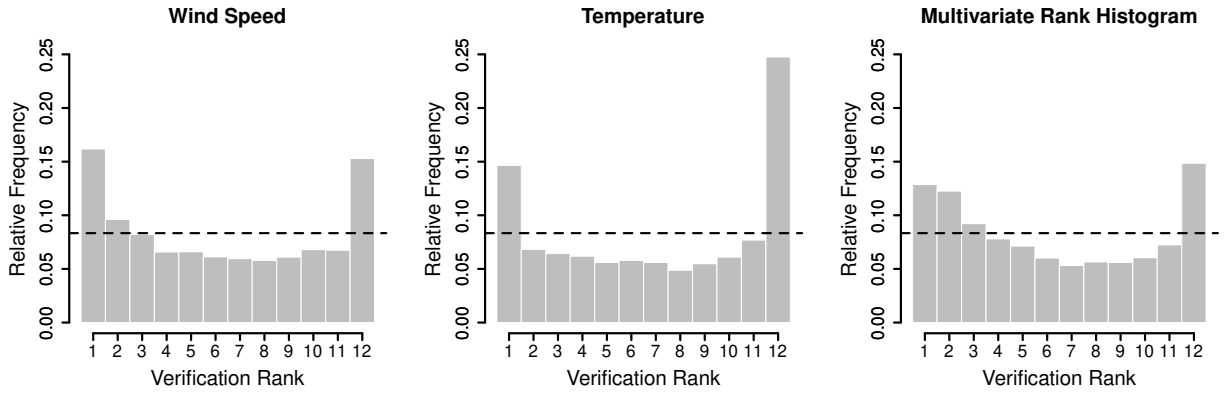
leads to the estimation of only one single correlation matrix over all considered stations instead of station specific correlation matrices.

5.4.1 Data

University of Washington mesoscale ensemble

Our study is based on 48 h UWME forecasts and corresponding validating observations of 10 m maximum wind speed (given in m/s, for a detailed description see Section 3.3.1) and 2 m minimum temperature (given in K) covering the same domain (Pacific Northwest of the United States) as in the case studies of Chapters 3 and 4. Again, we investigate only forecasts for calendar year 2008 with additional data from 2007 used for parameter estimation. After removing days and locations with missing data, 90 stations remain where the number of days for which forecasts and validating observations are available varies between 141 and 290.

Several studies have verified that wind speed and temperature forecasts of the UWME are strongly underdispersive (see e.g. Thorarinsdottir and Gneiting, 2010; Fraley *et al.*, 2010), and consequently uncalibrated. Obviously, the lack of calibration will remain valid if one considers these ensemble forecasts together, as predictions of a bivariate weather quantity. The underdispersive character of the raw ensemble can nicely be observed in Figure 5.1 displaying the univariate verification rank histograms of wind speed and temperature forecasts together with their joint multivariate rank histogram. The corresponding reliability indices Δ defined by (1.4.1) are 0.647, 0.842 and 0.550, respectively, and in many cases the raw ensemble either over-, or underestimates the verifying observation. Further, the need of bivariate modelling can be justified both by the positive correlation of 0.125 of the verifying observations of wind speed and temperature for calendar year 2008 taken along all dates and locations, and by the correlations of 0.187 and 0.189 of forecast errors of the ensemble median and mean, respectively.



(a) (b) (c)

Figure 5.2: Verification rank histograms of the ALADIN-HUNEPS ensemble forecasts of wind speed (a) and temperature (b) and the multivariate rank histogram (c) for the period 1 April 2012 – 31 March 2013.

ALADIN-HUNEPS ensemble

Besides the 11-member wind speed ensemble forecasts and observations described in detail in Section 3.3.1, in this study we consider the matching ensembles of 42 h forecasts of 2 m temperature (given in K) produced by the ALADIN-HUNEPS EPS, together with the corresponding validating observations for the one-year period between 1 April 2012 and 31 March 2013 and for the period from 1 October 2010 to 25 March 2011. For more details about the investigated ALADIN-HUNEPS temperature data sets see Baran *et al.* (2014a,b).

ALADIN-HUNEPS wind speed and temperature forecasts are better calibrated than those of the UWME, however, the rank histograms in Figure 5.2 still exhibit a strong underdispersive character. The bivariate reliability index equals 0.317, whereas the reliability indices of wind speed and temperature are 0.322 and 0.455, respectively. The need of bivariate post-processing is again supported by the forecast error correlations of 0.119 and 0.123 of the ensemble median and mean, respectively, however, in this case the verifying observations of wind speed and temperature show a very slight negative correlation of -0.029. This latter difference compared with the UWME, where this correlation equals 0.125, might be explained by the different types of wind and temperature quantities being examined (maximal vs. instantaneous).

5.4.2 Verification results for the UWME

In the present case study we apply the same training period length of 40 days as in Möller *et al.* (2013) which was determined with the help of an exploratory data analysis on a subset of the data set. Since in our rolling training periods for estimating the BMA and EMOS parameters we can also use data from calendar year 2007, predictive distributions can be produced for the whole calendar year 2008. This means 291 calendar days (after excluding dates with missing data) and a total of 24 302 individual forecast cases. As the eight ensemble members of the UWME are not exchangeable, for calibration we apply bivariate BMA models (5.1.2) and (5.1.3) and EMOS model (5.2.1) with $M = 8$.

	Probabilistic forecasts			Median forecasts			Mean forecasts		
	ES	Δ	DS	EE	ϱ	ϱ_{err}	EE	ϱ	ϱ_{err}
BMA	2.110	0.015	2.250	2.973	0.154	0.182	2.972	0.155	0.183
Pars. BMA	2.117	0.033	2.286	2.967	0.180	0.182	2.967	0.171	0.182
Indep. BMA	2.124	0.048	2.320	2.977	0.163	0.175	2.977	0.151	0.177
Copula BMA	2.089	0.030	2.272	2.977	0.160	0.176	2.978	0.152	0.177
EMOS	2.127	0.025	2.273	2.982	0.165	0.182	2.982	0.157	0.182
Indep. EMOS	2.118	0.059	2.206	2.966	0.164	0.176	2.966	0.155	0.178
Copula EMOS	2.088	0.021	2.169	2.967	0.162	0.178	2.967	0.156	0.179
Raw ensemble	2.562	0.550	0.773	3.087	0.017	0.187	3.072	0.007	0.189

Table 5.1: Mean energy score (ES), reliability index (Δ) and mean determinant sharpness (DS) of probabilistic forecasts, mean Euclidean error (EE) of point forecasts (median/mean), empirical correlation (ϱ) and empirical correlation of errors (ϱ_{err}) of wind speed and temperature components of point forecasts for the UWME. Empirical correlation of observations corresponding to the forecast cases: 0.125.

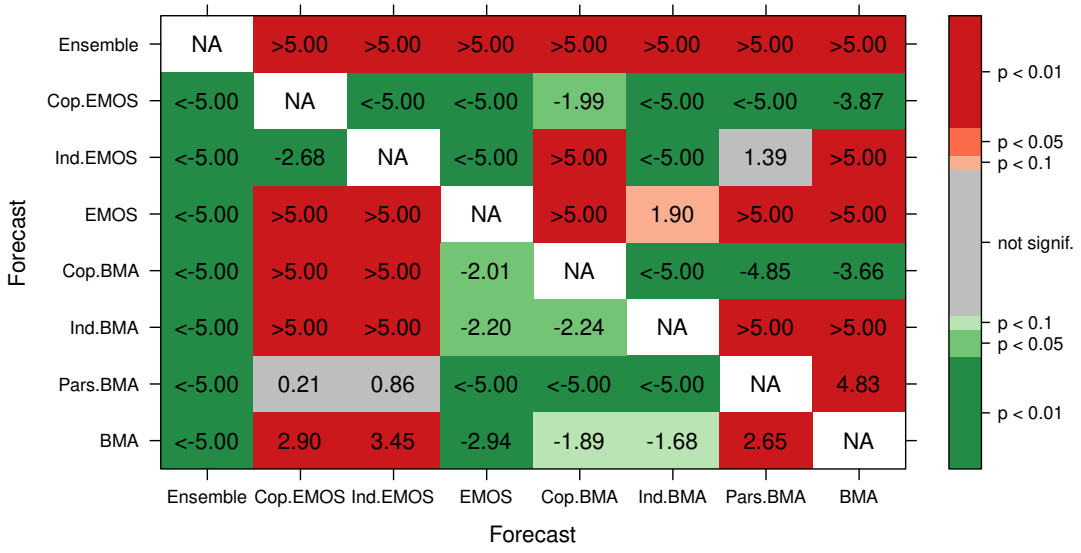


Figure 5.3: Values of the test statistic of the two-tailed DM test for equal predictive performance based on ES (*upper triangle*) and EE of median forecasts (*lower triangle*) for the UWME data. Green/red entries indicate superior performance of the forecast in the corresponding row/column.

In the case of the copula method, data from calendar year 2007 are applied for estimating the correlation between the two weather quantities, and the resulting correlation matrix is then employed for the analysis of the 2008 data.

In Table 5.1 the verification scores calculated using the BMA model (5.1.2) and its parsimonious version (5.1.3), the EMOS model (5.2.1), the independent BMA and EMOS models of wind speed and temperature, the copula model of Möller *et al.* (2013) both

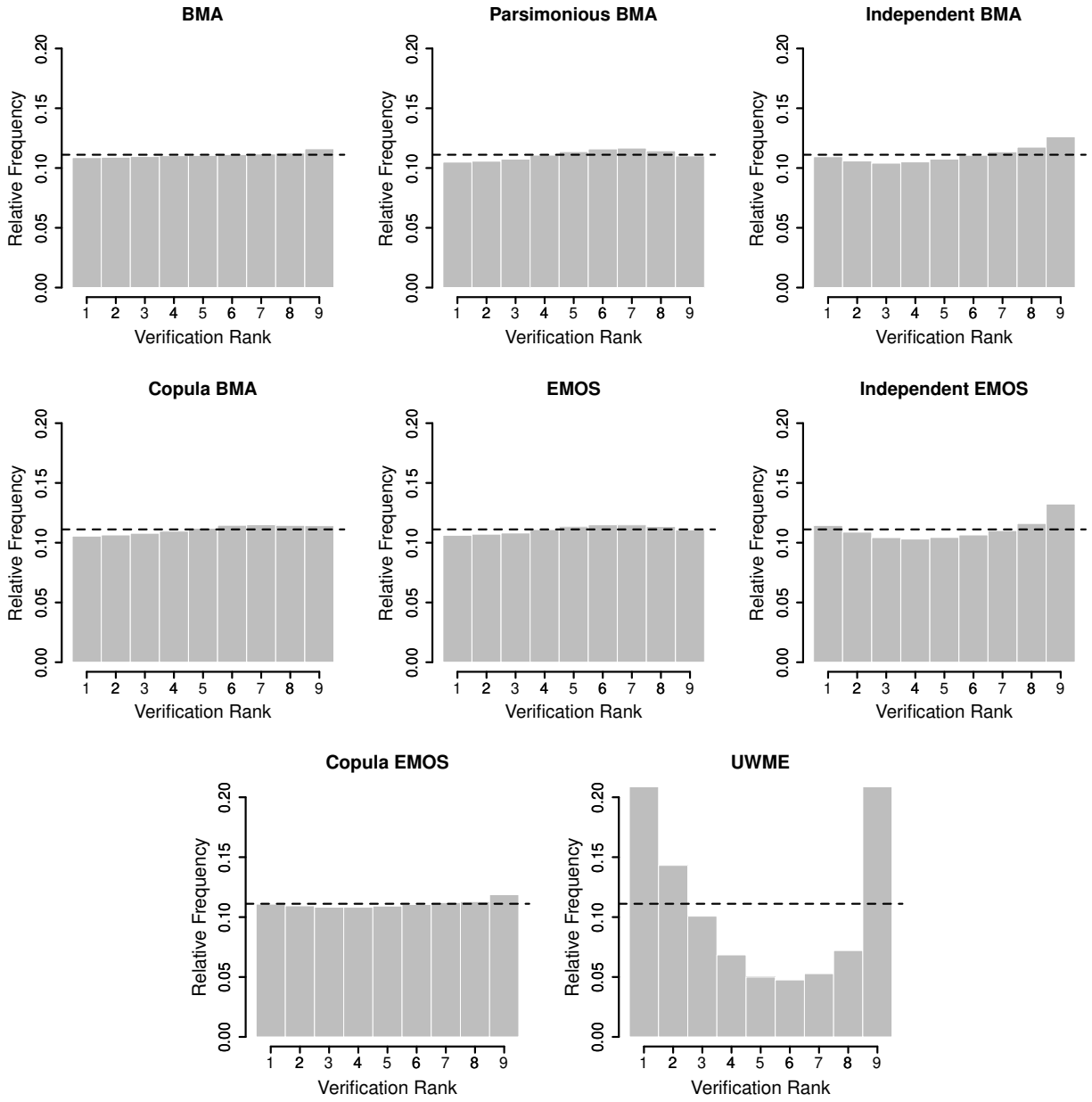


Figure 5.4: Multivariate rank histograms of BMA, parsimonious BMA, independent BMA, BMA based Gaussian copula, EMOS, independent EMOS, EMOS based Gaussian copula post-processed and raw UWME forecasts of maximum wind speed and minimum temperature. Average p-values of chi-square tests for uniformity (mean significance for 10000 random samples of sizes 2500 each): BMA: 0.439; parsimonious BMA: 0.304; independent BMA: 0.132; BMA based Gaussian copula: 0.347; EMOS: 0.373; independent EMOS: 0.046; EMOS based Gaussian copula: 0.382.

with BMA and EMOS post-processed marginals and the raw ensemble are given, whereas Figure 5.3 contains the results of two-tailed Diebold-Mariano tests of equal predictive performance in terms of the mean energy score (ES) and mean Euclidean error (EE) of median forecasts. Compared with the raw ensemble, all post-processing techniques substantially improve the calibration of probabilistic forecasts, which is quantified by

the significant decrease of the ES and large change in the reliability index (Δ). The improvement can also be observed in Figure 5.4 showing the corresponding multivariate rank histograms based either on samples from the various predictive distributions or on raw ensemble forecasts. Although the chi-square test rejects uniformity for all post-processing models, the mean p-values of 10000 random samples of multivariate ranks of sizes 2500 each nicely reflect the shapes of the corresponding histograms of Figure 5.4 and reliability indices Δ of Table 5.1. The price to pay for the better calibration is the substantial loss in sharpness (see the corresponding values of DS), however, this is a direct consequence of the small dispersion of the raw ensemble (see Figure 5.1). Post-processing also results in slightly (but significantly) smaller mean Euclidean errors (EE) indicating more accurate median and mean forecasts. Further, the empirical correlations ϱ of the wind and temperature components of the post-processed point forecasts are much closer to the correlation of 0.125 of the verifying observations than the corresponding correlations of the ensemble median and mean which are smaller by a magnitude. This latter is a weakness of the raw ensemble, however, one should also remark that all error correlations ϱ_{err} (including the raw ensemble) are very similar to each other (around 0.180).

Comparing the different post-processing techniques it is noticeable that the main difference between the various approaches to calibration appears in the reliability index. The bivariate BMA model results in the smallest Δ value, followed by the EMOS based Gaussian copula and the bivariate EMOS methods, which is in line with shapes of the corresponding multivariate rank histograms plotted in Figure 5.4. The large Δ values and the slightly U-shaped rank histograms of the independent BMA and EMOS approaches support the idea of bivariate modelling. Further, BMA model (5.1.2) outperforms its parsimonious counterpart (5.1.3) in terms of ES, Δ and DS, whereas the smallest energy scores correspond to the two copula approaches. However, in the model choice one should also take into account that copula methods require additional data for estimating the correlation matrix, whereas in the BMA and EMOS approaches the parameters are estimated using only the training data. Finally, in case of the latter two methods the computational costs (see Section 5.4.4) might also have an influence on the decision.

5.4.3 Verification results for the ALADIN-HUNEPS ensemble

In the case of the ALADIN-HUNEPS ensemble we consider the same natural grouping of ensemble members into two groups as in Sections 3.3 and 4.3.3. The first group contains just the control member $\boldsymbol{f}_c$, whereas in the second are the 10 statistically indistinguishable ensemble members $\boldsymbol{f}_{p,1}, \ldots, \boldsymbol{f}_{p,10}$, initialized from randomly perturbed initial conditions. This leads us to the BMA predictive PDF

$$p\big(\boldsymbol{x}|\, \boldsymbol{f}_c\, \boldsymbol{f}_{p,1}, \ldots, \boldsymbol{f}_{p,10}; A_c, A_p; B_c, B_p; \Sigma\big) = \omega g\big(\boldsymbol{x}|\, A_c + B_c \boldsymbol{f}_c, \Sigma\big) + \frac{1-\omega}{10} \sum_{\ell=1}^{10} g\big(\boldsymbol{x}|\, A_p + B_p \boldsymbol{f}_{p,\ell}, \Sigma\big), \tag{5.4.1}$$

which is a particular case of model (5.1.2), and to its parsimonious version

$$q\big(\boldsymbol{x}|\, \boldsymbol{f}_c\, \boldsymbol{f}_{p,1}, \ldots, \boldsymbol{f}_{p,10}; A, B; \Sigma\big) = \omega g\big(\boldsymbol{x}|\, A + B\boldsymbol{f}_c, \Sigma\big) + \frac{1-\omega}{10} \sum_{\ell=1}^{10} g\big(\boldsymbol{x}|\, A + B\boldsymbol{f}_{p,\ell}, \Sigma\big) \tag{5.4.2}$$

	Probabilistic forecasts			Median forecasts			Mean forecasts		
	ES	Δ	DS	EE	ϱ	ϱ_{err}	EE	ϱ	ϱ_{err}
BMA	1.434	0.031	1.539	2.004	−0.032	0.129	2.007	−0.041	0.129
Pars. BMA	1.428	0.021	1.534	1.999	−0.031	0.131	1.998	−0.035	0.128
Indep. BMA	1.454	0.015	1.573	2.033	−0.018	0.119	2.032	−0.030	0.119
Copula BMA	1.393	0.063	1.526	2.032	−0.021	0.119	2.031	−0.030	0.119
EMOS	1.442	0.034	1.478	2.015	−0.041	0.132	2.016	−0.049	0.132
Indep. EMOS	1.436	0.051	1.456	2.002	−0.033	0.128	2.002	−0.044	0.127
Copula EMOS	1.384	0.075	1.557	2.000	−0.036	0.128	2.000	−0.044	0.127
Raw ensemble	1.623	0.327	0.935	2.102	−0.068	0.122	2.083	−0.060	0.124

Table 5.2: Mean energy score (ES), reliability index (Δ) and mean determinant sharpness (DS) of probabilistic forecasts, mean Euclidean error (EE) of point forecasts (median/mean), empirical correlation (ϱ) and empirical correlation of errors (ϱ_{err}) of wind speed and temperature components of point forecasts for the ALADIN-HUNEPS ensemble. Empirical correlation of observations corresponding to the forecast cases: −0.033.

corresponding to model (5.1.3), where $\omega \in [0, 1]$, and g is defined by (5.1.1). The bivariate EMOS predictive distribution can be given as a special case of model (5.2.2), namely

$$\mathcal{N}_2^0\big(\mathcal{A} + \mathcal{B}_c \boldsymbol{f}_c + \mathcal{B}_p \overline{\boldsymbol{f}}_p, \mathcal{C} + \mathcal{D}\boldsymbol{S}\mathcal{D}^\top\big),$$

where $\overline{\boldsymbol{f}}_p$ is the mean vector of the 10 exchangeable ensemble members.

Based on a preliminary data analysis (univariate BMA and EMOS calibration of wind speed and temperature forecasts) we use a 40 days training period. In this way ensemble members, validating observations and BMA and EMOS models are available for the period 12 May 2012 – 31 March 2013 (just after the first 40 days training period having 318 calendar days, since on six days all ensemble members are missing). In line with the case study performed in Möller *et al.* (2013), additional data of the period 1 October 2010 – 25 March 2011 are utilized to estimate the correlation matrices of the Gaussian copula models. For the BMA and EMOS fits that are employed to estimate the correlation structure, a 40 days training period was used as well. The resulting (global) correlation matrices are then carried forward into the analysis of the 2012/2013 data.

The effects of statistical calibration of ensemble forecasts are quantified by the multivariate scores reported in Table 5.2, whereas Figure 5.5 gives the results of DM tests for equal predictive performance for energy score and Euclidean error of median forecasts. Compared with the raw ensemble all seven post-processing methods result in significantly lower energy scores and substantially smaller reliability indices (see also Figure 5.6). Similar to the UWME, one can also observe a significant loss in determinant sharpness which is again an effect of the underdispersive nature of the ensemble. However, here the increase in DS is around 60 %, whereas for the UWME the raw ensemble is almost three times sharper than the various predictive PDFs. This again indicates the better calibration of the ALADIN-HUNEPS ensemble, which is fully consistent with Figures 5.1 and 5.2 and the corresponding reliability indices given in Section 5.4.1. Further, the ensemble median and mean vectors produce significantly larger Euclidean errors than the corresponding post-processed point forecasts. Moreover, the empirical correlations of the components of

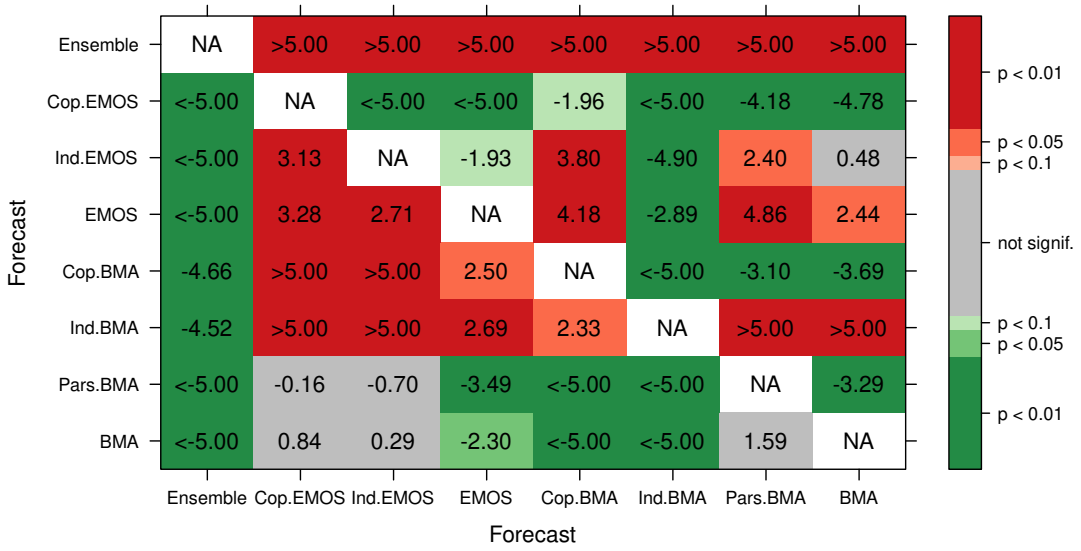


Figure 5.5: Values of the test statistic of the two-tailed DM test for equal predictive performance based on ES (*upper triangle*) and EE of median forecasts (*lower triangle*) for the ALADIN-HUNEPS data. Green/red entries indicate superior performance of the forecast in the corresponding row/column.

the ensemble median and mean are almost the double of the nominal correlation -0.033 of observations, whereas the correlations of wind speed and temperature components of the BMA and EMOS point forecasts are close to this value. Finally, both the ensemble median/mean and their calibrated counterparts exhibit almost the same forecast error correlations.

From the competing post-processing methods the Gaussian copula approach with EMOS marginals results in the lowest energy score and the second lowest Euclidean errors, however, the differences compared with the corresponding scores of the bivariate BMA and EMOS models (especially in the EE values) are rather small, and in some cases not significant (see Figure 5.5). Reliability indices show far larger variability and the highest scores belong to the two copula models followed by the independent EMOS approach. The Δ values in Table 5.2 are in accordance with the corresponding rank histograms in Figure 5.6 and mean p-values of chi-square tests for uniformity as well: the rank histograms of both copula methods are strongly hump-shaped indicating overdispersion, whereas the histogram of the independent EMOS approach exhibits some underdispersion. For the ALADIN-HUNEPS ensemble the two bivariate BMA models have the best overall performance, closely followed by the bivariate EMOS method, however, similar to the case of the UWME, the computational costs might also effect the model choice.

5.4.4 Computational aspects

As mentioned in Section 3.3.2, in the case of BMA calibration the bottleneck with respect to the computation costs is the EM algorithm applied for ML estimation of the parameters. Although the bivariate BMA model described is Section 5.1 makes use of a modification

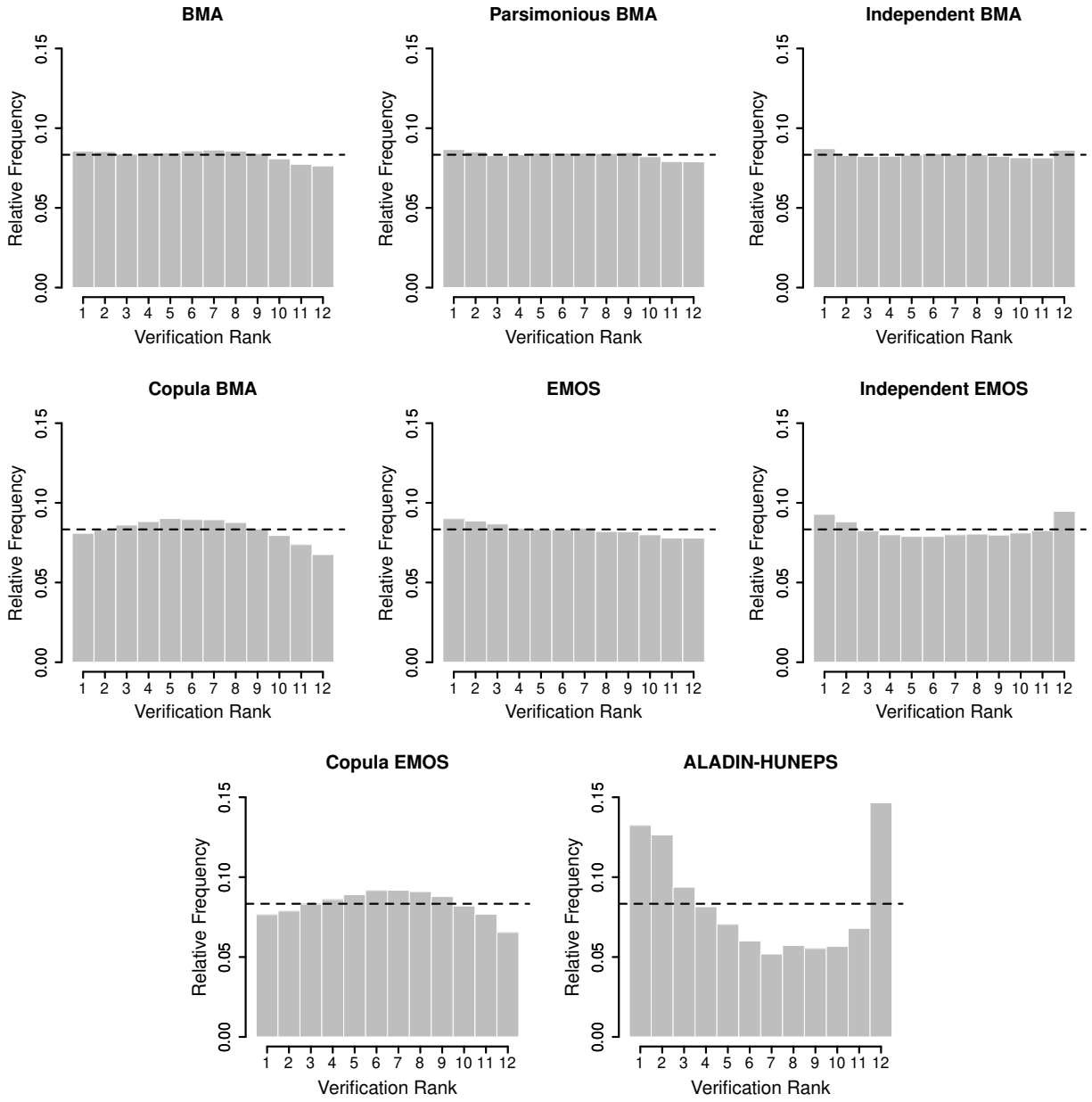
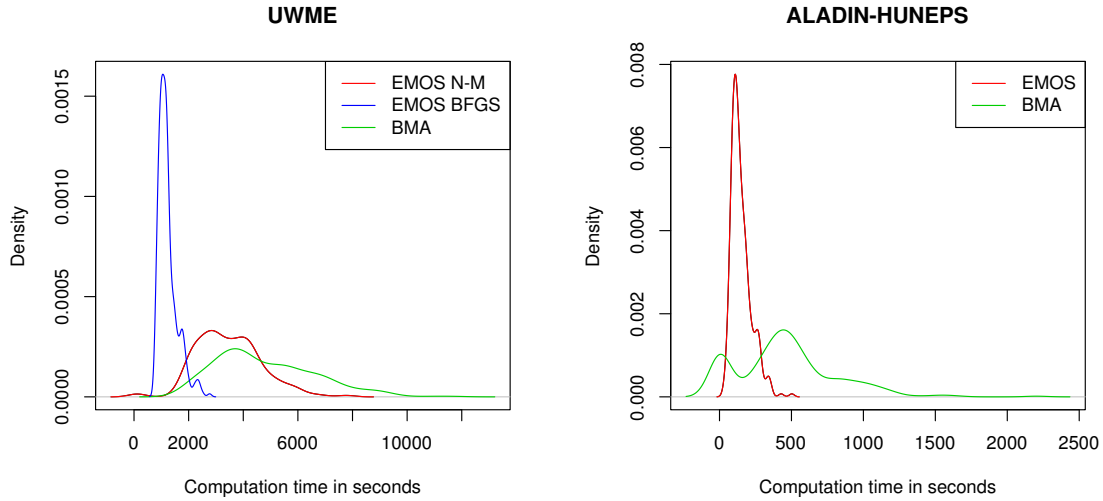


Figure 5.6: Multivariate rank histograms of BMA, parsimonious BMA, independent BMA, BMA based Gaussian copula, EMOS, independent EMOS, EMOS based Gaussian copula post-processed and raw ALADIN-HUNEPS ensemble forecasts of wind speed and temperature for the period 12 May 2012 – 31 March 2013. Average p-values of chi-square tests for uniformity (mean significance for 10000 random samples of sizes 2500 each): BMA: 0.320; parsimonious BMA: 0.408; independent BMA: 0.449; BMA based Gaussian copula: 0.046; EMOS: 0.278; independent EMOS: 0.132; EMOS based Gaussian copula: 0.019.

of the truncated data EM algorithm for Gaussian mixture models (Lee and Scott, 2012) which operates with closed formulae and there is no need of numerical optimization in the M step, due to the large number of free parameters (UWME: 59; ALADIN-HUNEPS: 17) it requires quite a lot of iterations resulting in long computation times.

For the EMOS methods the most time-consuming and problematic part of ensemble



(a) (b)

Figure 5.7: Densities of computation times for the bivariate BMA and EMOS models. (a) UWME for the calendar year 2008; (b) ALADIN-HUNEPS ensemble for the period 12 May 2012 – 31 March 2013.

post-processing is the numerical optimization used in parameter estimation. In the case of bivariate EMOS calibration of the ALADIN-HUNEPS ensemble only the robust Nelder-Mead algorithm occurs to be reliable, as one has to estimate 18 free parameters with the help of 400 forecast cases of the training data. For the UWME the data/parameter ratio is much better, as 42 free parameters have to be estimated using on average 3354 forecast cases. For this data set the reported Nelder-Mead and the faster BFGS algorithm give almost the same results.

Figures 5.7a and 5.7b show the kernel density estimates of the distribution of computation times over the days in the verification period for bivariate BMA and EMOS models (implemented in `R`) for the UWME and ALADIN-HUNEPS ensemble, respectively, calculated again on the same portable computer as in the previous case studies (Intel Quad Core i7-4700MQ CPU (2.40GHz $\times$ 4), 20 Gb RAM). We remark that in Figure 5.7a the density of computation times of the EMOS model with BFGS optimization is also plotted. The densities displayed in Figure 5.7 clearly show that in terms of computation time the EMOS model outperforms the BMA approach. However, one should also remark that these computation times are still too long for an operational use.

Finally, the Gaussian copula method starts with fast univariate BMA or EMOS calibrations, where the mean computation times allocated to parameter estimation of e.g. wind speed/temperature EMOS models for individual days in the verification periods of the UWME and the ALADIN-HUNEPS ensemble are 2.193/4.908 and 0.097/0.068 seconds, respectively. However, this approach utilizes an additional data set for estimating the correlation matrix of the Gaussian copula on the basis of additional post-processing of the univariate predictive PDFs. Hence, in terms of computational efficiency the presented version of the copula method is not comparable with the bivariate approaches and it is excluded from our analysis. Note that a model estimating EMOS parameters and copula covariances dynamically from the training data would be more appropriate for comparison. Such a dynamic approach of estimating the copula correlation was briefly investigated for

the case study of Möller *et al.* (2013), but did not yield significant improvement over the static approach.

5.5 Conclusions

In this chapter bivariate BMA and EMOS models for joint calibration of ensemble forecasts of wind speed and temperature are described which are based on a bivariate normal distribution truncated from below at zero in its first coordinate. The model is tested on wind speed and temperature forecasts of the 8-member University of Washington mesoscale ensemble and of the 11-member ALADIN-HUNEPS ensemble of the Hungarian Meteorological Service. These ensemble prediction systems differ both in the weather quantities being forecast and in the generation of the ensemble members.

Using appropriate verification measures (energy score, reliability index and determinant sharpness of probabilistic forecasts and Euclidean errors, correlations, as well as correlations of errors of median/mean forecasts) the predictive performance of the bivariate models is compared with the forecast skills of the independent BMA and EMOS calibrations of wind speed and temperature, the Gaussian copula method of Möller *et al.* (2013) based on both univariate BMA and univariate EMOS models and the raw ensemble vectors as well.

From the results of the presented case studies one can conclude that compared with the raw ensemble, post-processing always improves the calibration of probabilistic and accuracy of point forecasts. In terms of predictive performance the bivariate models are able to keep up with the more general Gaussian copula approach, however, without requiring an additional data set for estimating the correlations. Further, concerning the computational costs, bivariate EMOS approach outperforms the bivariate BMA calibration.

Finally, one should remark that the Gaussian copula approach can be applied for any desired type and number of weather quantities, whereas the current versions of the bivariate BMA and EMOS models are applicable only for a bivariate weather quantity vector where the components can be assumed to be normal and truncated normal. However, such low dimensional parametric post-processing methods can serve as building blocks e.g. for non-parametric calibration approaches taking into account spatial dependence (Schefzik, 2016b).

Chapter 6

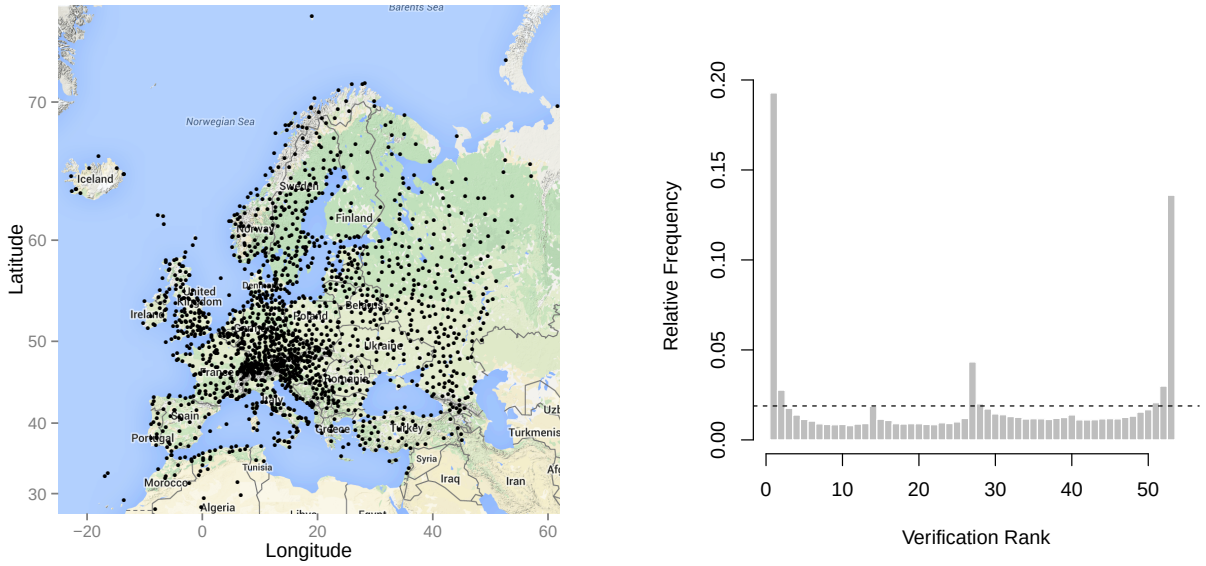
Semi-local approaches to parameter estimation

As it has been discussed in Section 1.3.3, for selecting the training sets for parameter estimation in BMA and EMOS modelling two basic approaches are given by local and regional methods. In the local approach, only forecast cases from the single observation station of interest are considered for the parameter estimation, whereas in the regional approach, data from all available observation stations are composited to form a single training set for all stations. Local estimation generally results in better predictive performance, however, numerically it is often problematic if only limited amounts of training data are available. In contrast, there are typically no numerical stability issues in regional parameter estimation, however, in the case of large ensemble domains it is undesirable to obtain a single set of coefficients for all observation stations due to the potentially significant differences in the climatological properties of the observation stations and forecast errors of the ensemble.

In this chapter we apply the truncated normal EMOS model of Thorarinsdottir and Gneiting (2010) described in Section 3.2.1 for statistical post-processing of wind speed forecasts of the 52-member Grand Limited Area Model Ensemble Prediction System (GLAMEPS; Iversen *et al.*, 2011). The GLAMEPS ensemble covers a large domain across Europe and Northern Africa, however, only a short period of data is available. This makes both regional and local estimation problematic. Two similarity-based semi-local approaches to parameter estimation are described (Lerch and Baran, 2017) in order to account for these challenges. The distance-based approach uses data from stations with similar characteristics to augment the training data for a given station and follows ideas of Hamill *et al.* (2008), whereas the clustering-based approach employs k-means clustering to obtain groups of similar observation stations with respect to various features which then form shared training sets for parameter estimation within each cluster.

6.1 The GLAMEPS ensemble

The GLAMEPS ensemble is a short-range multi-model EPS launched in 2006 as a part of the cooperation between the Aire Limitée Adaptation dynamique Developpement International (ALADIN) and High Resolution Limited Area Modelling (HIRLAM) consortia. It operates on a large domain covering Europe, North-Africa and the Northern Atlantic and



(a) (b)

Figure 6.1: Locations of observation stations (a) and verification rank histogram (b) of the GLAMEPS ensemble.

the currently running Version 2 (GLAMEPSv2) is a combination of the subensembles from two versions of the Aire Limitée Application de la Recherche à l'Operationnel (ALARO) model (intéractions soil biosphere atmosphère (ISBA) and surface externalisée (SURFEX) schemes, see e.g. Noilhan and Planton (1989) and Hamdi *et al.* (2014)) and two versions of the HIRLAM model (Kain-Fritsch and soft transition condensation (STRACO) schemes, see e.g. Kain and Fritsch (1990) and Sass (2002)). Each subensemble consists of 12 perturbed members and a control forecast, and half of the perturbed members are lagged by 6 h (Deckmyn, 2014).

Our data base contains 52 ensemble members of 18 h ahead forecasts of 10 m wind speed for 1 738 observation sites (see Figure 6.1a) together with the corresponding validating observations for 2 October – 25 November 2013, and 2 February – 18 May 2014. We divide the available data into two equally large periods from October 2013 to February 2014 and from March 2014 to May 2014 in order to allow for rolling training periods of sufficient length. The forecasts are evaluated over the second period. Data from the first period are used to obtain training periods of equal lengths for all days, and to determine the similarities between the stations, see Section 6.2.2 for details.

The U-shaped verification rank histogram of the GLAMEPS ensemble depicted in Figure 6.1b indicates that the GLAMEPS wind speed forecasts lack calibration and are underdispersive, i.e. too many observations fall outside the ensemble range. This deficiency can be observed for various ensemble prediction systems, see e.g. Baran and Lerch (2015) or the case studies of Chapter 3.

6.2 EMOS models for the GLAMEPS ensemble

To calibrate GLAMEPS ensemble forecasts we apply the truncated normal EMOS model (3.2.5), where location and scale parameters are affine functions of the ensemble means of the different exchangeable groups of ensemble members and the ensemble variance, respectively.

6.2.1 Model formulations

The link functions connecting the parameters of the predictive distribution of the EMOS models and the ensemble forecasts depend on the stochastic properties of the ensemble. The GLAMEPS ensemble consists of four subensembles which differ in the choice of numerical model and parametrization scheme. Each subensemble contains a control and $6 + 6$ (non-lagged and lagged) perturbed members. This induces a natural grouping into twelve groups:

- ALARO model with ISBA parametrization scheme, group of size 6 with group mean $\overline{f}_{AI}$;
- ALARO model with SURFEX parametrization scheme, group of size 6 with group mean $\overline{f}_{AS}$;
- HIRLAM model with Kain-Fritsch parametrization scheme, group of size 6 with group mean $\overline{f}_{HK}$;
- HIRLAM model with STRACO parametrization scheme, group of size 6 with group mean $\overline{f}_{HS}$;
- lagged versions of above groups, 4 individual groups of size 6 with group means $\overline{f}_{\bullet L}$, where $\bullet \in \{AI, AS, HK, HS\}$;
- control forecasts $f_{AI,c}, f_{AS,c}, f_{HK,c}, f_{HS,c}$, 4 individual groups of size 1.

The members within each individual group are exchangeable and should share a common set of EMOS coefficients, resulting in a predictive TN distribution with location

$$\begin{aligned} a_0 + a_{AI,c} f_{AI,c} + \big(a_{AI}\overline{f}_{AI} + a_{AIL}\overline{f}_{AIL}\big) + a_{AS,c} f_{AS,c} + \big(a_{AS}\overline{f}_{AS} + a_{ASL}\overline{f}_{ASL}\big) \\ + a_{HK,c} f_{HK,c} + \big(a_{HK}\overline{f}_{HK} + a_{HKL}\overline{f}_{HKL}\big) + a_{HS,c} f_{HS,c} + \big(a_{HS}\overline{f}_{HS} + a_{HSL}\overline{f}_{HSL}\big) \end{aligned} \tag{6.2.1}$$

and scale $b_0 + b_1 S^2$, which is a special case of model (3.2.5). This model has a total number of 15 parameters to be estimated and will be referred to as *full model*.

A natural simplification is to assign the same parameter values to the lagged and non-lagged exchangeable ensemble members of a subensemble, which results in a reduced model with location

$$\begin{aligned} a_0 + a_{AI,c} f_{AI,c} + a_{AI}\big(\overline{f}_{AI} + \overline{f}_{AIL}\big) + a_{AS,c} f_{AS,c} + a_{AS}\big(\overline{f}_{AS} + \overline{f}_{ASL}\big) \\ + a_{HK,c} f_{HK,c} + a_{HK}\big(\overline{f}_{HK} + \overline{f}_{HKL}\big) + a_{HS,c} f_{HS,c} + a_{HS}\big(\overline{f}_{HS} + \overline{f}_{HSL}\big) \end{aligned} \tag{6.2.2}$$

and 11 parameters to be estimated. This model will be referred to as *lag ignoring model*.

Finally, we also investigate the fully exchangeable situation where the existence of the aforementioned groups is ignored, and all ensemble members are assumed to form a single exchangeable group. In this case the predictive distribution is given by

$$\mathcal{N}_0^\infty\big(a_0 + a_1\overline{f}, b_0 + b_1 S^2\big), \tag{6.2.3}$$

where again, $\overline{f}$ denotes the ensemble mean, and we refer to this model as *simplified model*.

6.2.2 Similarity-based semi-local parameter estimation

In general, the coefficients of the TN EMOS model are estimated by minimizing the mean CRPS of the predictive distributions over suitably chosen rolling training periods consisting of the preceding n days. As described in Section 1.3.3, there exist two basic approaches for selecting the training data: regional and local. The regional (or global) approach composites ensemble forecasts and validating observations from all available stations during the rolling training period. In the case of the GLAMEPS ensemble regional estimation of parameters means that a single set of coefficients is used for the wide-ranging domain and the geographical and climatological variability might thus not be sufficiently taken into account. Although this approach can be implemented without numerical stability issues and offers slight gains in predictive performance compared with the raw ensemble (see Section 6.3), there is room for further improvement for large and heterogeneous domains.

By contrast, the local approach produces distinct parameter estimates for different stations by using only the training data of the given station. Local models typically result in better predictive performance compared with regional models (see e.g. Thorarinsdottir and Gneiting, 2010; Schuhen *et al.*, 2012), however, these training sets contain only one observation per day and the estimation of local EMOS models thus requires significantly longer training periods to avoid numerical stability issues. For example, in the case of the GLAMEPS data, full model (6.2.1) has 15 parameters to be estimated, which makes the use of local EMOS problematic.

We propose two alternative similarity-based semi-local approaches which avoid the problems that make both regional and local estimation of the EMOS coefficients undesirable for the GLAMEPS data. The basic idea of the semi-local methods is to combine the advantages of regional and local estimation by augmenting the training data for a given station with data from stations with similar characteristics. The choice of similar stations is either based on suitably defined distance functions or on clustering.

Distance-based semi-local model

Following Hamill *et al.* (2008), the training sets of a given station are increased by including training data from other stations with similar features. The similarity between stations is determined based on suitably defined distance functions. We use the term *distance function* in a general sense with only one of the proposed similarity measures depending on the actual geographical locations of the observation stations. From a mathematical point of view, all considered distance functions are semi-metrics, i.e. non-negative and symmetric functions $d : \{1, \ldots, 1738\} \times \{1, \ldots, 1738\} \to \mathbb{R}$ with $d(i, i) = 0$. Distance functions can thus be seen as negatively oriented similarity measures with smaller values indicating more similar characteristics of the stations of interest. Note that compared with Hamill

et al. (2008), we consider alternative choices of distance functions, and our forecasts are evaluated over a set of observation stations whereas the forecasts and analysis data used by Hamill *et al.* (2008) are given on a grid where different conclusions may apply.

Generally, the distance between two stations i and j denoted by $d(i,j)$ with $i,j \in \{1,\ldots,1738\}$ is determined using the first period of available data from October 2013 to February 2014 which is distinct from the verification period. In the semi-local estimation of the EMOS model for a given station i_0, we then add the corresponding forecast cases in the rolling training period from the L most similar stations, i.e. the L stations with the smallest distances $d(i_0,j)$, $j \in \{1,\ldots,1738\}$.

Alternatively, one could also iteratively determine the similarities anew in every rolling training period. However, this approach requires lots of computational resources as the $\frac{1737\cdot 1738}{2} \approx 1.5 \times 10^6$ pairwise distances between stations have to be re-computed for every training period, and is thus infeasible due to the large number of observation stations. In particular, note that already the non-iterative semi-local model estimation with a fixed set of similarities is computationally more demanding compared with local parameter estimation which arises as a special case for $L = 1$. Furthermore, initial tests did not indicate substantial improvements in the predictive performance for the GLAMEPS data, we thus limit our discussion to the use of a fixed period of data for determining the similarities.

We investigate the following four distance functions.

Distance 1: Geographical locations. The distance between stations i and j is given by the Euclidean distance of the locations $(\mathcal{X}_i, \mathcal{Y}_i)$ and $(\mathcal{X}_j, \mathcal{Y}_j)$ of the two stations, i.e.

$$d^{(1)}(i,j) := \sqrt{(\mathcal{X}_i - \mathcal{X}_j)^2 + (\mathcal{Y}_i - \mathcal{Y}_j)^2}.$$

The Euclidean distance is employed here since the station locations in the data set are given on the linearly transformed model estimation grid. In general, the great circle distance is a more appropriate distance measure for actual geographical locations on the globe.

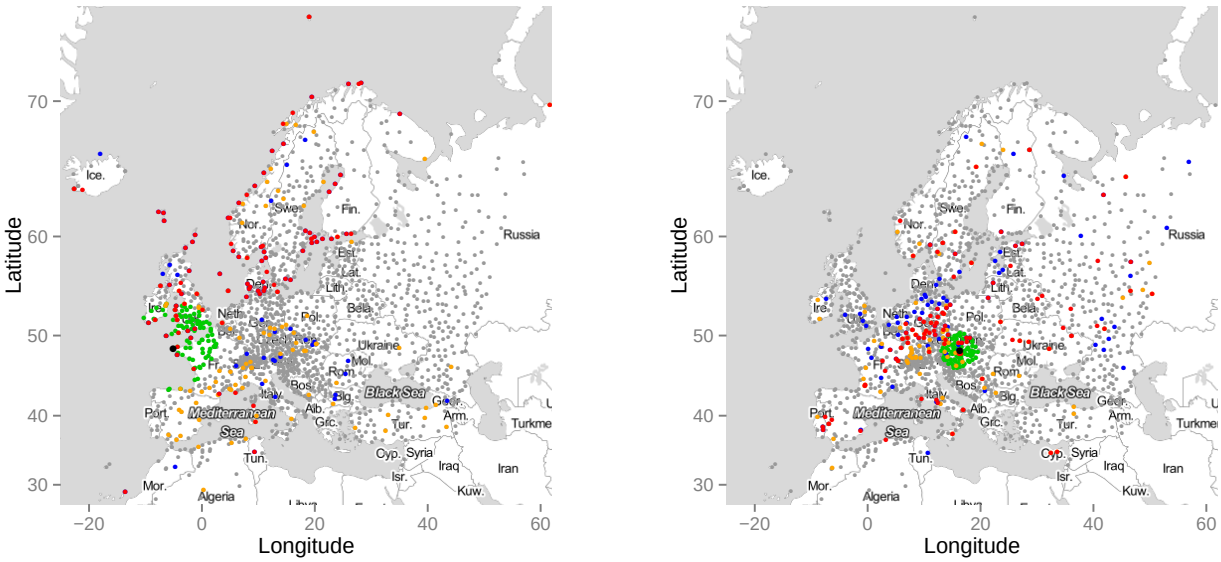
Distance 2: Station climatology. Let $\widehat{F}_i$ denote the empirical CDF of wind speed observations at station i over the first period of data. Similarly to the distance function which was proposed by Hamill *et al.* (2008), the distance to station j is given by the normalized sum over the absolute differences of the respective empirical CDFs $\widehat{F}_i$ and $\widehat{F}_j$ evaluated at a set of fixed values S, i.e.

$$d^{(2)}(i,j) := \frac{1}{|S|} \sum_{x \in S} \left| \widehat{F}_i(x) - \widehat{F}_j(x) \right|,$$

where $|S|$ denotes the cardinality of S. Here, we choose $S = \{0, 0.5, 1, 1.5, \ldots, 14.5, 15\}$ (equidistant evaluation points between the minimum observation of 0 m/s and the 99th percentile of all observations at 15 m/s) and note that the obtained sets of similar stations are somewhat robust to minor changes in the definition of the set of evaluation points, e.g. setting $S = \{0, 1, \ldots, 20\}$ results in very similar sets of close stations.

Distance 3: Ensemble forecast errors. Denote the ensemble mean for station i and date t by $\overline{f}_{i,t}$ and the corresponding verifying observation by $x_{i,t}$, then the forecast error $e_{i,t}$ of the ensemble mean is given by

$$e_{i,t} = \overline{f}_{i,t} - x_{i,t}.$$



(a) (b)

Figure 6.2: Illustration of the 100 most similar stations measured by the four distance functions for two reference stations at Ouessant, France (a) and Vienna, Austria (b). The reference stations are indicated by black dots. Several points are part of the set of similar stations in more than one similarity measure; in this case they are assigned the color of the last mentioned distance.

The third distance function is based on the distribution of these forecast errors. To that end, we define the empirical CDF of the forecast errors at station i as

$$\widehat{G}_i^e(z) := \frac{1}{|T|} \sum_{t \in T} \mathbb{I}_{\{\overline{f}_{i,t} - x_{i,t} \leq z\}}, \tag{6.2.4}$$

where T denotes the set of dates in the first period of data. The distance between two stations i and j is then given by

$$d^{(3)}(i,j) := \frac{1}{|S'|} \sum_{x \in S'} \left| \widehat{G}_i^e(x) - \widehat{G}_j^e(x) \right|,$$

where $S' = \{-10, -9.5, -9, -8.5, \ldots, 0, \ldots, 8.5, 9, 9.5, 10\}$ denotes the set of fixed values at which the empirical CDFs of the forecast errors are evaluated.

Distance 4: Combination of distances 2 and 3. We add up the values of distances 2 and 3 to define a distance function which depends on both the climatology of the observations as well as the distribution of the forecast errors of the ensemble. With the above notation the corresponding distance function is

$$d^{(4)}(i,j) := d^{(2)}(i,j) + d^{(3)}(i,j) = \frac{1}{|S|} \sum_{x \in S} \left| \widehat{F}_i(x) - \widehat{F}_j(x) \right| + \frac{1}{|S'|} \sum_{x \in \tilde{S}'} \left| \widehat{G}_i^e(x) - \widehat{G}_j^e(x) \right|.$$

Figure 6.2 illustrates the four distance functions for two of the observation stations by displaying the 100 most similar stations in a specific colour each. For the station at Ouessant (Figure 6.2a) located on the North-Western coast of France, it can be observed that the 100 most similar stations measured by the distance functions depending on the distribution of observations and forecast errors (distances 2–4) are mostly located at coastal regions and islands in Northern Europe, in particular if these characteristics are combined (distance 4). By contrast, the most similar stations to the observation site at Vienna (Figure 6.2b) are distributed over continental central Europe, mostly located in France, Germany and Poland. Because of the differences in the density of the observation station network, the stations in close geographical proximity to the reference station at Ouessant are spread out over larger geographical distances compared with the respective stations around Vienna. Therefore, data from stations with different climatological properties might be added to the training sets for parameter estimation which indicates a potential drawback of the location-based distance 1.

Clustering-based semi-local model

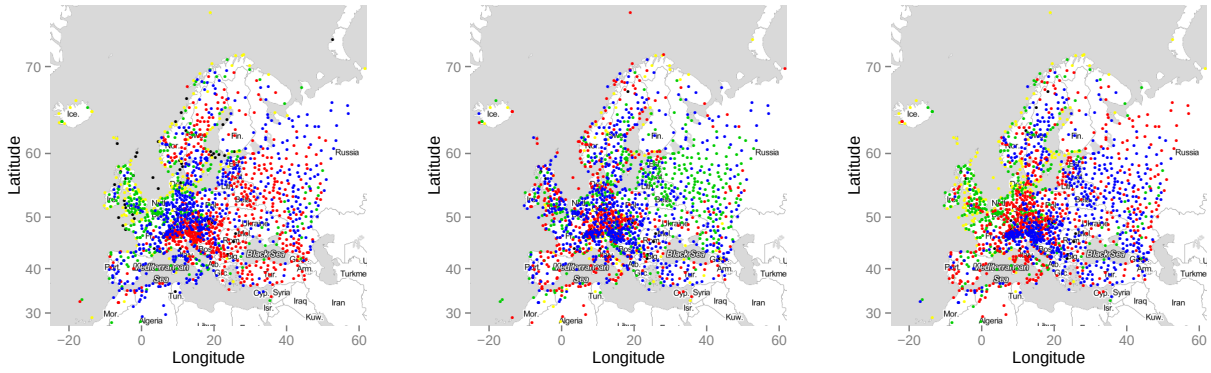
Further, as an alternative to the distance-based approach we propose a semi-local approach based on cluster analysis. Here, the observation sites are grouped into clusters, and parameter estimation is performed for each cluster individually using only ensemble forecasts and validating observations at stations within the given cluster. To determine the clusters of observation stations we apply k-means clustering (see e.g. Hastie *et al.*, 2009) to various choices of feature sets which are based on climatological characteristics of the observation stations and the distribution of forecast errors.

In comparison with the distance-based method, the clustering-based semi-local approach is computationally much more efficient as the parameter estimation is only performed for k distinct training sets for each given day, whereas the distance-based approach requires individual estimation of the coefficients at each of the 1738 stations with partially overlapping training sets. Further, the similarities between the observation stations are obtained in a more efficient way as clustering is computationally less demanding compared with the computation of pair-wise distances between all observation stations (up to symmetry). In particular, clustering-based semi-local estimation is also computationally more efficient than local parameter estimation which arises as a special case with $k = 1738$ clusters of size 1 each. In this light, clustering-based semi-local models offer a compromise between adaptivity and parsimony of the numerical estimation.

The discussion does not account for the computational costs of the actual clustering. However, there are efficient algorithms for k-means clustering, e.g. the Hartigan-Wong algorithm (Hartigan and Wong, 1979), which converge rapidly for the data at hand. The costs of the actual clustering are thus negligible compared with the computational costs of the numerical parameter estimation. In contrast with the distance-based approach, this allows for iteratively determining the clusters anew in every training period without a significant increase in the overall computational costs. This adaptive approach will be pursued for all clustering-based semi-local models discussed below.

We denote the number of features used in the k-means clustering procedure by N and consider the following feature sets.

Feature set 1: Station climatology. Let $\widehat{F}_{i,n}$ denote the empirical CDF of the wind speed observations at station i over the rolling training period consisting of the preceding



(a) (b) (c)

Figure 6.3: Illustration of cluster memberships of the observation stations based on feature sets (a) 1 (climatology), (b) 2 (forecast errors) and (c) 3 (climatology and forecast errors) obtained with a fixed number of 5 clusters and 24 features. Colours are assigned to the clusters by size (in descending order: blue, red, green, yellow, black).

n forecast cases at this station. The feature set for station i is given by the set of equidistant quantiles of $\widehat{F}_{i,n}$ at levels $\frac{1}{N+1}, \frac{2}{N+1}, \ldots, \frac{N}{N+1}$.

Feature set 2: Forecast errors. Denote the empirical CDF (6.2.4) of forecast errors $e_{i,t}$ by $\widehat{G}^e_{i,n}(z)$, where the set T in the expression $t \in T$ denotes the preceding n dates as the clusters are iteratively determined anew in every rolling training period. The feature set for station i is then given by the set of equidistant quantiles of $\widehat{G}^e_{i,n}$ at levels $\frac{1}{N+1}, \frac{2}{N+1}, \ldots, \frac{N}{N+1}$.

Feature set 3: Combination of feature sets 1 and 2. To define a feature set that depends on both the station climatology and the distribution of forecast errors, we combine equidistant quantiles of $\widehat{F}_{i,n}$ at levels $\frac{1}{N_1+1}, \ldots, \frac{N_1}{N_1+1}$ and equidistant quantiles of $\widehat{G}^e_{i,n}$ at levels $\frac{1}{N_2+1}, \ldots, \frac{N_2}{N_2+1}$ into one single set of size $N = N_1 + N_2$, where N_1 and N_2 are defined as follows. If N is an even number, let $N_1 = N_2 = \frac{N}{2}$, otherwise let $N_1 = \lceil \frac{N}{2} \rceil$ and $N_2 = N - N_1$.

Alternative choices of feature sets where the geographical location of the observation stations is included in the definition have also been investigated, but result in a reduction of the predictive performance and are thus omitted in the following discussion.

Figure 6.3 illustrates the clusters obtained for observation stations for the different feature sets with a fixed number of $k = 5$ clusters. For the feature set defined in terms of the distribution of the observations (feature set 1, Figure 6.3a), one can observe two larger clusters distributed over central Europe, where one cluster mainly contains stations in Germany and France, whereas the other contains most of the stations in the Alps and continental Eastern Europe. The remaining clusters are predominantly centred near the United Kingdom and coastal regions of France and Northern Europe. If the clusters are determined on the basis of forecast errors (feature set 2, Figure 6.3b), the stations are mainly grouped into three almost equally large clusters, where the most notable difference compared with the first feature set is the predominant presence of the third cluster in North-Eastern Europe. Further, the stations in the United Kingdom and coastal regions

of Europe now mostly belong to the two biggest clusters rather than forming separate sets. Clustering based on a combination of the distribution of the observations and forecast errors (feature set 3, Figure 6.3c) results in a pattern of cluster memberships in between the other two choices. In particular, the alpine regions, continental Europe and the coastal regions and the United Kingdom show the most clear-cut separation compared with the other feature sets.

6.3 Results

The forecast skill of the semi-local approaches of parameter estimation is tested on full, lag ignoring and simplified TN EMOS models for calibrating GLAMEPS wind speed ensemble forecasts (3.2.5) with location parameters linked to the ensemble via (6.2.1), (6.2.2) and (6.2.3), respectively. The results are compared with the predictive performance of regional and local approaches.

6.3.1 Selection of tuning parameters for semi-local parameter estimation methods

Both semi-local parameter estimation techniques require the choice of various tuning parameters given by the length of the rolling training period, the number of similar stations to be taken into account, the number of features and the number of clusters. We now discuss the effect of these tuning parameters on the predictive performance of the forecast models. For that, the full, lag ignoring and simplified model were estimated using the distance-based and clustering-based semi-local parameter estimation techniques described in Section 6.2.2. Conclusions are drawn based on the mean CRPS over the evaluation period. For comparison, note that the average CRPS values of the GLAMEPS ensemble and the best regional TN model are 1.058 and 0.955, respectively. Because of numerical stability issues in the parameter estimation, a comparison to local models is impossible, an estimate of the mean CRPS of the locally estimated simplified TN model (6.2.3) can be obtained if problematic parameter estimates (around 0.1 % of the forecast cases) are replaced by corresponding estimates from preceding forecast cases. The mean CRPS of the local simplified model with such subsequent modifications equals 0.790, see Section 6.3.2.

Distance-based approach

In the distance-based semi-local approach to parameter estimation, the size of the training set for a given station i is increased by including corresponding training data from the L most similar stations, i.e. the L stations with the smallest distances $d(i, j)$, $j \in \{1, 2, \ldots, 1738\}$. Note that for the distance functions defined in Section 6.2.2, $d(i, i) = 0$, a value of, e.g. $L = 5$ thus means that the training set for station i consists of data from this station, and of data from the 4 stations with the smallest distances to station i. Figure 6.4 illustrates the effect of L on the predictive performance measured as mean CRPS of the three proposed models for selected lengths of the training period.

For distance 1 the predictive performance decreases with the number of similar stations added to the training sets, except for the more complex lag ignoring and full models

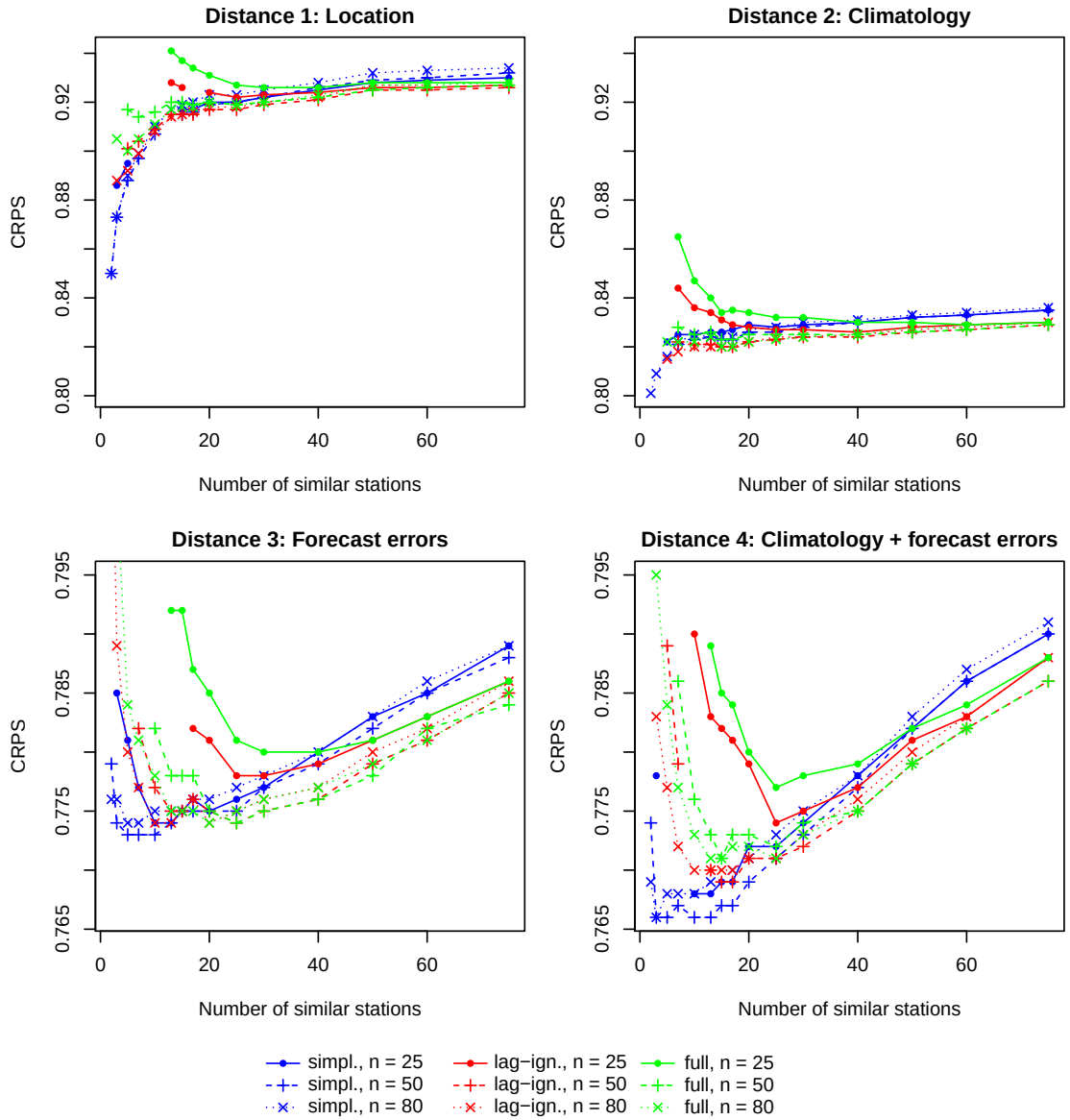


Figure 6.4: Effect of the number of similar stations L on the predictive performance of the distance-based semi-local models for three choices of training period lengths n (in days) (missing line segments indicate unsuccessful parameter estimation for these choices of tuning parameters; note the different scales of the plots in the first and second row caused by the varying predictive performances of the respective models).

and shorter training periods, where the best CRPS values are attained for values around $L = 20$. Clearly, the inclusion of similar stations then allows for unproblematic parameter estimation, but generally, if the similarities are determined based on geographical locations, as few stations as possible should be used in order to achieve results as close as possible to the favourable (but even for long training periods impossible) local parameter estimation corresponding to $L = 1$. Similar conclusions apply for the climatology-based distance 2, however, the predictive performance of these models is notably better.

A different pattern emerges for distances 3 and 4 based on forecast errors and combinations with climatology shown in the second row of Figure 6.4. In contrast with distances

1 and 2, augmenting the training sets with data from similar stations here generally improves the forecasts. The best predictive performances are achieved with choices of L between 10 and 30 depending on the similarity measure and the length of the training period, whereas smaller values of L result in worse predictions. The mean CRPS increases for values of L exceeding around 30, however, note that these semi-local models still perform better than the local model for a wide range of tuning parameter values.

The effect of the length of the rolling training periods consisting of the preceding n days can also be seen from Figure 6.4 where each individual plot contains three different choices of n. Together with further investigations of plots of the average CRPS against the employed training period lengths (which are not shown), one can observe that n only has a small effect on the predictive performance of the models. For all considered distance functions, the forecast skill increases slightly with longer training periods, in particular for the more complex models and smaller values of L. This is to be expected from the smaller size of the training sets as parameter estimation becomes problematic for short training periods and few additional forecast cases from similar stations taken into account.

The simplified models show a slight decrease in predictive performance for training periods longer than 40–50 days, however, the differences are negligible compared with those between models based on varying choices of distance functions or varying numbers of similar stations taken into account. The overall best predictive performances across the three considered model formulations are achieved with training period lengths of 80 days.

Clustering-based approach

In the clustering-based semi-local approach k-means clustering based on the different feature sets is employed to group the observation stations into clusters. The lower computational costs of this approach allow for iterative computation of the clusters in every training period. This adaptive application of k-means clustering leads of improvements in mean CRPS of around $1 - 5\,\%$ compared with a non-iterative implementation.

Figure 6.5 illustrates the effect of the number of clusters k on the predictive performance. Choosing $k = 1$ obviously corresponds to regional parameter estimation. For all three feature sets considered here, the predictive performance increases for larger values of k up to around 100 clusters except for shorter training periods. Clearly, a larger number of clusters allows for a more refined grouping into sets of observation stations with similar characteristics. The predictive performance generally decreases if much more than $k = 100$ clusters are used. This behaviour is not surprising as the clusters become smaller and parameter estimation eventually becomes numerically unstable, particularly for the lag ignoring and full models. Depending on training period length and feature set, only small improvements can be observed for k exceeding values of around 40–70 clusters.

As observed for the distance-based models, the clustering-based semi-local models defined in terms of the distribution of forecast errors and the station climatology (feature sets 2 and 3) are able to outperform the local model over a wide range of tuning parameter choices except for short training periods. The worse predictive performance for shorter training periods is to be expected as the smaller amount of forecasts cases used to determine the clusters might result in a less accurate partitioning of the observation stations. Compared with the distance-based approach it can be observed that, for some

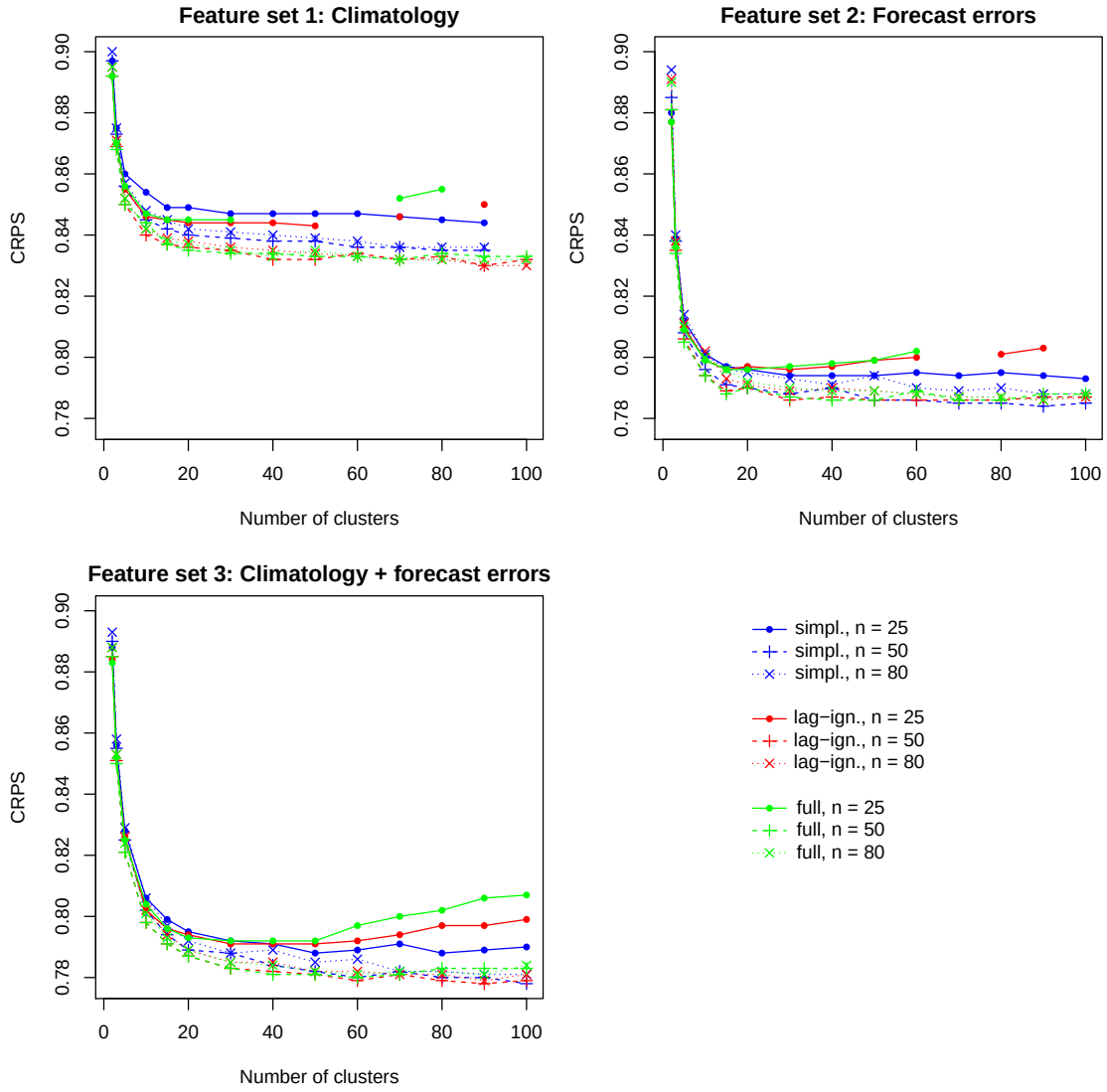


Figure 6.5: Effect of the number of clusters k on the predictive performance of clustering-based semi-local models for three choices of training period lengths n (in days) (all models are estimated with feature sets of size $N = 24$; missing line segments indicate unsuccessful parameter estimation for these choices of tuning parameters).

k, training period lengths below 80 days are optimal. However, in comparison with the effect of different choices of feature sets the effect of the length of the training period is negligible.

Thus far, all clustering-based semi-local models shown in Figure 6.5 were estimated for a fixed feature set size of $N = 24$. Further investigations (which are not shown) indicate that the feature set size has only a small effect on the predictive performance compared with varying choices of k or n as long as sufficiently many features (around 5–10 depending on the other tuning parameters) are used. Reasons for this behaviour include the aforementioned robustness of the obtained cluster memberships with regards to N. The best results across all considered tuning parameter combinations are generally obtained for feature set sizes between 20 and 40 thus justifying our previous choice of

$N = 24$.

6.3.2 Forecast performance

The predictive performance of the semi-local models is evaluated by computing the mean CRPS of probabilistic forecasts, the mean absolute error of median forecasts and coverage and average width of nominal 96.2 % central prediction intervals (see Section 1.4) for the models considered. These scores evaluated over the verification period 1 March – 18 May 2014. We use the local climatological forecasts given by the observations at the corresponding station during the rolling training periods, the raw GLAMEPS ensemble predictions, and probabilistic forecasts by the regional TN model as benchmark models. Although locally estimated models are desirable, the estimation of these models is highly problematic for the GLAMEPS data due to the issues discussed earlier. Even for the simplified model (6.2.3) with a maximum training period length of 80 days, numerical issues occur in the local parameter estimation; e.g. some scale parameters are estimated to be 0. In this case the problematic parameter estimates are replaced by the preceding ones. Note that such subsequent adjustments are not necessary for the semi-local or regional models. Further, neither the full nor the lag ignoring local model can be successfully estimated as the employed numerical optimization algorithms fail to converge or produce numerical errors.

For brevity, we limit our discussion to the simplified and lag ignoring models. Figures 6.4 and 6.5 indicate that the full models generally result in slightly worse predictive performance compared with the lag ignoring ones, therefore the additional computational costs of taking into account the lagging in the subensembles are not justified. Different conclusions may apply for other ensemble prediction systems with lagged members.

With regards to the tuning parameters for the semi-local approaches, we employ a fixed training period length of 80 days, and use a fixed number of $N = 24$ features for k-means clustering to ensure comparability across the different models. For the individual distance-based and clustering-based semi-local models we then choose suitable values for the number of most similar stations L and the number of clusters k from Figures 6.4 and 6.5. While the chosen tuning parameter combinations might not be the overall optimal values for the individual models, the results hold for a wide range of tuning parameter choices as indicated by the sensitivity considerations in Section 6.3.1.

To determine the optimal tuning parameter values for a new data set we suggest to follow common practice from the extant literature on ensemble post-processing, and testing various combinations of parameter values, perhaps on a shorter initial test set. For the GLAMEPS ensemble, our analysis indicates that the most influential tuning parameters for the semi-local model estimation are the number of similar stations L, and the number of clusters k, respectively, see Section 6.3.1 for details.

Table 6.1 shows the mean CRPS, MAE of median values and coverage and average width of 96.2 % central prediction intervals for the models considered. The raw GLAMEPS ensemble predictions outperform the climatological forecasts and provide sharp prediction intervals, however, at the cost of being uncalibrated. Regional TN models are able to improve the calibration of the ensemble and result in around 10 % better mean CRPS values, however, the semi-local approaches significantly outperform the regional approaches for all considered models and tuning parameter choices, see also Figures 6.4 and 6.5.

Among the distance-based semi-local models, the best predictive performances are

Forecast		CRPS (m/s)	MAE (m/s)	Coverage (%)	Av. width (m/s)
Local climatology		1.127	1.580	96.6	7.96
GLAMEPS ensemble		1.058	1.376	67.1	3.50
Regional TN models					
Simplified		0.957	1.324	90.3	6.36
Lag ignoring		0.955	1.320	90.3	6.33
Local TN models (with subsequent modifications)					
Simplified		0.790	1.100	88.7	5.12
Distance-based semi-local TN models					
D1 simplified	$L = 3$	0.873	1.218	90.2	5.99
D1 lag ignoring	$L = 3$	0.887	1.236	89.2	5.71
D2 simplified	$L = 5$	0.816	1.136	90.0	5.61
D2 lag ignoring	$L = 5$	0.815	1.136	89.6	5.42
D3 simplified	$L = 5$	0.774	1.083	90.3	5.25
D3 lag ignoring	$L = 10$	0.774	1.083	90.2	5.21
D4 simplified	$L = 3$	0.766	1.069	89.9	5.16
D4 lag ignoring	$L = 10$	0.770	1.075	90.0	5.18
Clustering-based semi-local TN models					
C1 simplified	$k = 70$	0.836	1.162	89.8	5.68
C1 lag ignoring	$k = 70$	0.832	1.156	89.6	5.55
C2 simplified	$k = 70$	0.789	1.103	89.9	5.25
C2 lag ignoring	$k = 70$	0.787	1.099	89.8	5.22
C3 simplified	$k = 70$	0.782	1.091	89.7	5.19
C3 lag ignoring	$k = 70$	0.781	1.090	89.7	5.17

Table 6.1: Mean CRPS of probabilistic, MAE of median forecasts and coverage and average width of 96.2 % nominal central prediction intervals evaluated over the second period of data from March to May 2014. A training period length of 80 days is used for all models and the feature set size for the clustering-based models is fixed at $N = 24$.

achieved by distance functions 3 and 4 which utilize the climatological distribution and its combination with the distribution of the forecast errors, respectively. These semi-local models are also able to outperform the local model for a wide range of tuning parameter choices without requiring subsequent corrections and further allow for a successful estimation of the more complex lag ignoring and full semi-local models. Except for distance 2, the simplified model generally performs slightly better than the lag ignoring one, however, the differences are negligible compared with the differences between the varying model estimation approaches.

We obtain similar results for the clustering-based semi-local models which perform slightly worse compared with the corresponding distance-based models, however, still outperform the regional models and the local model if the clusters are determined based on forecast errors and station climatology. Here, the lag ignoring models show better predictive performances compared with the simplified models but, again, the differences are small compared with the influence of the choice of feature sets.

Formal statistical tests of equal predictive ability were performed to assess the signif-

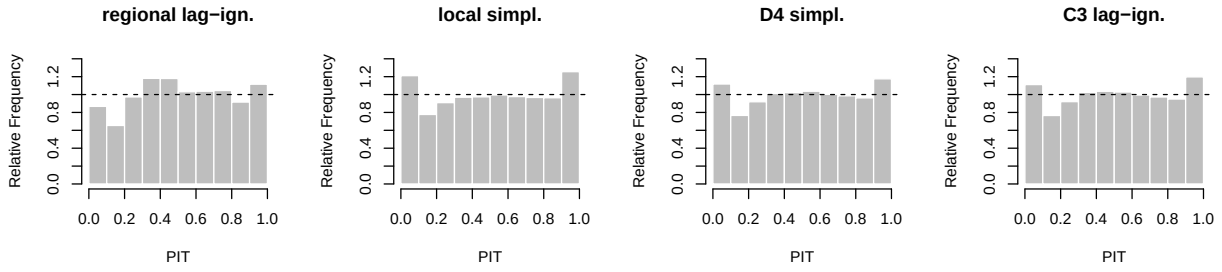


Figure 6.6: PIT histograms of the EMOS post-processed forecasts (all models are estimated with a rolling training period of 80 days; the semi-local models displayed are those with the best mean CRPS, see Table 6.1 for the corresponding tuning parameter choices).

icance of these findings. Two-sided Diebold-Mariano tests (see Section 1.4) based on the CRPS indicate that all observed score differences are significant at the 5 % level.

Figure 6.6 shows PIT histograms of the regional, the local and the semi-local models with the best mean CRPS values. Compared with the verification rank histogram of the raw GLAMEPS ensemble forecasts (see Figure 6.1b), all post-processing models exhibit substantially improved calibration with PIT histograms showing much smaller deviations from the desired uniform distribution. The hump-shaped PIT histogram of the regional TN model indicates a slight underprediction of lower wind speed values. The local and semi-local models are able to correct for this deficiency and show slightly better calibration, in particular for the semi-local models. However, all models consistently show a slight underdispersion that can also be seen from the coverage values reported in Table 6.1. This deficiency appears to be a general drawback of models based on the TN distribution (see the case studies of Chapter 3 or Baran and Lerch (2015)). Alternative distributional choices such as a weighted mixture of TN and log-normal distributions might lead to further improvements in calibration, as demonstrated in the case studies of Chapter 3 (see also Baran and Lerch, 2016). An application of the semi-local approach might be particularly interesting for this more complex model as many parameters have to be estimated and local estimation may thus not be feasible.

To conclude, we note that the overall best predictive performance is achieved by semi-local models where the similarities between stations are determined based on combinations of the distributions of observations and forecast errors at the given stations. Although all semi-local models show significantly better predictive performance than the regional models, these best models can also outperform the locally estimated model. The semi-local parameter estimation methods further allow for estimating more complex models without numerical issues, whereas local estimation is only possible for simplified model formulations with a reduced number of parameters and still requires subsequent modifications. Figures 6.4 and 6.5 indicate that these conclusions hold for a wide range of tuning parameter choices.

6.4 Conclusions

Two semi-local approaches to parameter estimation for ensemble post-processing are introduced where the training data for a given observation station are augmented with data

from stations with similar characteristics. The distance-based approach roughly follows the ideas of Hamill *et al.* (2008) and uses distance functions to determine the similarities between observations stations, whereas the novel clustering-based approach employs k-means clustering to obtain groups of similar stations.

The semi-local models outperform regional and local models and offer several advantages over these standard approaches to parameter estimation while being straightforward to implement. The clustering-based semi-local model estimation is further computationally much more efficient than local estimation. Although distance-based semi-local models show slightly better predictive performance compared with the clustering-based ones, the estimation requires substantially more computational resources. In particular, an iterative computation of the similarities in every training period is not feasible for the distance-based models. A recent application of the clustering-based semi-local approach can be found in Baran *et al.* (2019b), where the statistical post-processing of ECMWF global dual-resolution temperature ensemble forecasts is investigated.

Compared with the work of Hamill *et al.* (2008), several alternative distance functions are proposed and the distance-based approach for observations at specific stations is used instead of gridded data. It would be interesting to apply the novel similarity measures as well as the clustering-based approach to grid-based forecast and analysis data and assess potential differences. In particular, similarity measures incorporating the distribution of forecast errors (distances 3 and 4) might also offer improvements over the climatology-based distance function used by Hamill *et al.* (2008) when applied to gridded data. For this reason we have recently started experiments with clustering-based semi-local EMOS calibration of ECMWF gridded dual-resolution precipitation forecasts for Europe with the data set identical to the one studied in Gascón *et al.* (2019), where the authors select training data for non-parametric calibration using the approach of Hamill *et al.* (2008). Further, in connected works, Kleiber *et al.* (2011); Scheuerer and Büermann (2014), and Scheuerer and Möller (2015) consider alternative approaches incorporating techniques from geostatistics and novel model formulations that entail local adaptivity of the parameters, and allow for extrapolating the forecasts to locations or grid points without observations. These schemes are particularly important for interpolating local forecasts obtained at observation stations to the model grid.

The distance functions considered here are defined in terms of station locations, observations, and forecast errors of the ensemble. Alternative choices of similarity measures proposed in related works may lead to further improvements for different EPSs. For example, Schefzik (2016a) proposes a similarity measure defined in terms of mean and variance of the ensemble forecasts, and Kleiber *et al.* (2011) include covariates such as elevation and land use information. However, for the GLAMEPS predictions, similarities defined by characteristics of the ensemble overlap with location-based similarities (distance 1) to a great extent, and covariate information was not available for the data at hand.

The group memberships of the observation stations in the clustering-based semi-local models are determined by k-means clustering. Alternative clustering methods might potentially lead to improvements (for reviews and comparisons see e.g. Fraley and Raftery, 1998). We did not incorporate information on the geographical locations of the stations or characteristics of the ensemble into the selected feature sets as initial tests indicated a worse predictive performance. For different ensemble prediction systems, alternative choices of feature sets may lead to further improvements. One should also take into account that for a small number of observation stations with rather different climatology,

for some sites the proposed k-means clustering approach might result in local parameter estimation. This is the situation in the case study of Díaz *et al.* (2019) on calibration of ensemble forecasts of temperature, where the forecast domain contains the high-mountain region around Santiago de Chile. To estimate mean and variance parameters of the corresponding EMOS model (1.3.3) using a rather short training period of 20 days, instead of the proposed dynamic clustering, fixed clusters based on station elevation are preferred.

Further, Junk *et al.* (2015) propose analog-based local EMOS models where the training set for a given station is chosen by selecting forecast cases with similar ensemble forecasts for that station and similar ideas appear in Hemri and Klein (2017). This analog-based approach thus utilizes information for a given station in an optimal way by selecting subsets of the local training sets, whereas our semi-local models combine information from multiple observation stations based on similarities. Although the analog-based modification of the local parameter estimation method shows good predictive performance in case studies on hub height wind speed (Junk *et al.*, 2015) and water level (Hemri and Klein, 2017), it requires sufficiently long training periods for locally selecting similar forecast cases. The implementation of this analog-based approach is thus infeasible for the GLAMEPS data, however, comparisons and combinations with the similarity-based semi-local approaches proposed here are of interest and might result in further improvement in predictive performance.

Conclusions and discussion

We hope that the results presented in Chapters 2 – 6, despite they focus on the work of the author in probabilistic weather and hydrological forecasting, are able to provide an insight into the main parametric approaches to statistical post-processing of ensemble forecasts, give some heuristics behind the various models and also show the difficulties of model formulation, parameter estimation and forecast verification. For a very detailed summary of the current state of the art in statistical calibration we refer to the recent monograph of Vannitsem *et al.* (2018) containing all methods described here but the BMA model for post-processing water level ensemble forecasts introduced in Baran *et al.* (2019a) and explained in detail in Chapter 2.

Chapters 2 – 5 are formulated around different predictable quantities describing the corresponding BMA and EMOS models. As demonstrated, forecast skill of the raw ensemble is always significantly improved by statistical calibration, however, in most of the cases the is no unique winner, that is a method outperforming its competitors in all verification measures for all case studies. Thus, in order to choose the appropriate post-processing approach for the data set at hand, one has to look always at the whole picture including the calibration and sharpness of probabilistic forecasts, the accuracy of point forecasts, and in some situations also the computational costs.

The BMA model of Chapter 2 considerably outperforms the reference EMOS approach when classical rolling window training periods are used for estimating the model parameters. However, as our further tests indicate (see Baran *et al.*, 2019a), the use of analog-based selection of training periods from Hemri and Klein (2017) drastically decreases the gap in forecast skill, leaving only a small advantage of BMA compared with EMOS. This indicates that using a more sophisticated post-processing approach or the use of a smarter selection of training data are fairly redundant. Accordingly, we can recommend to use EMOS with analog-based training periods if a sufficiently long set of hydrological data is available and BMA otherwise. As an extension of the current study and direction of future research, following the ideas of Hemri *et al.* (2015) and Bellier *et al.* (2018), one can combine the BMA calibrated forecasts corresponding to different locations and lead times either into temporally, or both spatially and temporally coherent multivariate predictions with the help of modern techniques such as e.g. the ensemble copula coupling (Schefzik *et al.*, 2013) or the Gaussian copula approach (Pinson and Girard, 2012).

As demonstrated in Chapter 3, several different BMA and EMOS models exist for calibrating ensemble forecasts of wind speed. The proposed truncated normal BMA provides a significantly faster algorithm for estimating model parameters than the competing gamma BMA approach of Sloughter *et al.* (2010) and exhibits better forecast skill. The paper presenting this approach (Baran, 2014) is already highly cited and our next task is to have the model tested in many other case studies. We are planning to get the trun-

cated normal BMA approach included in the `ensembleBMA` package of `R`, which will make the method available to a wide range of users. Further, the EMOS approaches utilizing a log-normal distribution justified their raison d'etre by showing very good predictive performance in various case studies. Besides the situations investigated by the author, LN and TN-LN mixture EMOS models serve as benchmark models e.g. in investigating the forecast skill of the recent constrained quantile regression spline approach of Bremnes (2019). As mentioned in Section 3.4, the simplest direction of further research could be the use of a truncated GEV distribution in EMOS modelling, in order to correct the disadvantage of predicting negative wind speeds with positive probability of the currently available GEV distribution based approaches. More ambitious plans are the generalization of the time series model of Möller and Groß (2016) to non-Gaussian (e.g. truncated normal) variables and the modification of the Markovian EMOS approach of Möller *et al.* (2015) in order to incorporate spatial dependencies into wind speed modelling.

From the weather variables investigated in the present work, the calibration of precipitation forecasts occurs to be the most difficult task. On the one hand, the non-negative predictive distribution is not absolutely continuous, as it should put a positive weight on zero precipitation. On the other hand, precipitation data sets contain lots of zero observations, that is, large training sets are required for reliable parameter estimation in the case of positive precipitation amounts. The post-processing methods described in Chapter 4 show two different solutions to the first problem and in the presented two case studies the novel CSG EMOS model significantly outperforms the more complex gamma BMA approach and shows slightly better predictive performance than the GEV EMOS model. However, as pointed out by Hamill *et al.* (2017), "a method that has been demonstrated to produce high-quality post-processed guidance with a lengthy training data set and a single-model ensemble will not necessarily perform optimally with multi-model ensembles and short training data sets." Thus, a reasonable direction of further research is to investigate the calibration of precipitation accumulation forecasts using the semi-local approaches of Chapter 6, which allow the use of much shorter rolling training periods.

The bivariate BMA and EMOS models of Chapter 5 demonstrate that inter-variable dependencies of various weather quantities can be successfully modelled in a parametric way. In the presented case studies both proposed methods outperform the raw ensemble forecasts with a large margin and are able to keep up with the more general Gaussian copula approach of Möller *et al.* (2013). Obviously, parametric models tailored to specific weather variables are restricted to low-dimensional settings, however, as demonstrated by Schefzik (2016b), they can serve as components of more complex calibration methods assessing also spatial dependencies. Following this idea, one can also try to extend the above mentioned Markovian EMOS approach to obtain a spatial bivariate model for joint calibration of wind speed and temperature ensemble forecasts. A different direction of further studies is the application of low-dimensional parametric models for calibration of ensemble forecasts of user-oriented variables, which quantities have recently gained an increasing interest (see e.g. the past *Forecast Verification Metric Challenge*[1] of the Joint Working Group on Forecast Verification Research of World Meteorological Organization's World Weather Research Programme or the *TIGGE/S2S Challenge*[2] of the ECMWF).

[1] `http://www.wmo.int/pages/prog/arep/wwrp/new/FcstVerChallenge.html` [Accessed on 16 June 2019]

[2] `https://www.ecmwf.int/sites/default/files/medialibrary/2018-11/TIGGE-S2S-WS_Challenge.pdf` [Accessed on 16 June 2019]

As an example one can consider Thom's Discomfort Index (Stathopoulou *et al.*, 2005) combining temperature and relative humidity. From a reasonable joint predictive PDF of the latter two weather quantities one can obtain a predictive distribution for post-processing of ensemble forecasts of the derived Discomfort Index.

Finally, the distance-based and clustering-based semi-local approaches presented in Chapter 6 provide general tools to training data selection both for parametric and non-parametric post-processing methods. Besides the EMOS calibration of dual-resolution temperature forecasts with the help of the clustering-based method (Baran *et al.*, 2019b), the possibility of extending the state of the art post-processing approaches with semi-local parameter estimation is mentioned in several recently published works. As examples one can mention Holman *et al.* (2018), which investigates the calibration of wind vectors using a bivariate normal EMOS model, or van Straaten *et al.* (2018), where both non-parametric quantile regression forests and a parametric method based on a zero-adjusted gamma distribution are applied to post-processing of high-resolution ensemble precipitation forecasts.

Bibliography

Bao, L., Gneiting, T., Raftery, A. E., Grimit, E. P. and Guttorp, P. (2010) Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon. Weather Rev.* **138**, 1811–1821.

Baran, S. (2014) Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Comput. Stat. Data. Anal.* **75**, 227–238.

Baran, S., Hemri, S. and El Ayari, M. (2019a) Statistical post-processing of water level forecasts using Bayesian model averaging with doubly-truncated normal components. *Water Resour. Res.* **55**, 3997–4013.

Baran, S., Horányi, A. and Nemoda, D. (2013) Statistical post-processing of probabilistic wind speed forecasting in Hungary. *Meteorol. Z.* **22**, 273–282.

Baran, S., Horányi, A. and Nemoda, D. (2014a) Probabilistic temperature forecasting with statistical calibration in Hungary. *Meteorol. Atmos. Phys.* **124**, 129–142.

Baran, S., Horányi, A. and Nemoda, D. (2014b) Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Időjárás* **118**, 217–241.

Baran, S. and Lerch, S. (2015) Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Q. J. R. Meteorol. Soc.* **141**, 2289–2299.

Baran, S. and Lerch, S. (2016) Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics* **27**, 116–130.

Baran, S. and Lerch, S. (2018) Combining predictive distributions for statistical post-processing of ensemble forecasts. *Int. J. Forecast.* **34**, 477–496.

Baran, S., Leutbecher, M., Szabó, M. and Ben Bouallègue, Z. (2019b) Statistical post-processing of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.* **145**, 1705–1720.

Baran, S. and Möller, A. (2015) Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics* **26**, 120–132.

Baran, S. and Möller, A. (2017) Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteorol. Atmos. Phys.* **129**, 99–112.

Baran, S. and Nemoda, D. (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics* **27**, 280–292.

Bassetti, F., Casarin, R. and Ravazzolo, F. (2018) Bayesian nonparametric calibration and combination of predictive distributions. *J. Am. Stat. Assoc.* **113**, 675–685.

Bellier, J., Zin, I. and Bontron, G. (2018) Generating coherent ensemble forecasts after hydrological postprocessing: adaptations of ECC-based methods. *Water Resour. Res.* **54**, 5741–5762.

Ben Bouallègue, Z., Theis, S. E. and Gebhardt, C. (2013) Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorol. Z.* **22**, 49–59.

Bentzien, S. and Friederichs, P. (2012) Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather Forecast.* **27**, 988–1002.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.-Y., Parsons, D., Raoult, B., Schuster, D., Dias, P. S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L. and Worley, S. (2010) The THORPEX interactive grand global ensemble. *B. Am. Meteorol. Soc.* **91**, 1059–1072.

Böhning, D. (2014) The 2nd special issue on advances in mixture models. *Comput. Stat. Data. Anal.* **71**, 1–2.

Bremnes, J. B. (2019) Constrained quantile regression splines for ensemble postprocessing. *Mon. Weather Rev.* **147**, 1769–1780.

Broc̆ker, J. and Smith, L. A. (2007) Increasing the reliability of reliability diagrams. *Weather Forecast.* **22**, 651–661.

Buizza, R. (2018) Ensemble forecasting and the need for calibration. In Vannitsem, S., Wilks, D. S., Messner, J. W. (eds.), *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, Amsterdam, pp. 15–48.

Buizza, R., Tribbia, J., Molteni, F. and Palmer T. (1993) Computation of optimal unstable structures for a numerical weather prediction system. *Tellus A* **45**, 388–407.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**, 1076–1097.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16**, 1190–1208.

Chen, S. C. and Lindsay, B. (2014) Improving mixture tree construction using better EM algorithms. *Comput. Stat. Data. Anal.* **74**, 17–25.

Cloke, H. L. and Pappenberger, F. (2009) Ensemble flood forecasting: A review. *J. Hydrol.* **375**, 613–626.

Deckmyn, A. (2014) Introducing GLAMEPSv2. *ALADIN Forecasters Meeting,* Ankara, Turkey, 10–11 September, 2014. Available at: `http://www.cnrm.meteo.fr/aladin/meshtml/FM2014/presentation/AladinFm_AD_be.pdf` [Accessed on 16 June 2019]

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X. and Stull, R. B. (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.* **111** D24307, doi:10.1029/2005JD006917.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39**, 1–39.

Dennis, J. and Schnabel, R. (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, New Jersey.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P. and Cébron, P. (2015) PEARP, the Météo-France short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **141**, 1671–1685.

Díaz, M., Nicolis, O., Marín, J. C. and **Baran, S.** (2019) Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile. *Meteorol. Appl.*, under review.

Diebold, F. X., Gunther, T. and Tay, A. (1998) Evaluating density forecasts, with applications to financial risk management. *Int. Econ. Rev.* **39**, 863–883.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253–263.

Duan, Q., Ajami, N. K., Gao, X. and Sorooshian, S. (2007) Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**, 1371–1386.

Eckel, F. A. and Mass, C. F. (2005) Effective mesoscale, short-range ensemble forecasting. *Weather Forecast.* **20**, 328–350.

ECMWF Directorate (2012) Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter* **133**, 11–13.

Fraley, C. and Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.

Fraley, C., Raftery, A. E. and Gneiting, T. (2010) Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Weather Rev.* **138**, 190–202.

Fraley, C., Raftery, A. E., Gneiting, T., Sloughter, J. M. and Berrocal, V. J. (2011) Probabilistic weather forecasting in R. *R J.* **3**, 55–63.

Friederichs, P. and Thorarinsdottir, T. L. (2012) Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* **23**, 579–594.

Fritz, H., Filzmoser, P. and Croux, C. (2012) A comparison of algorithms for the multivariate L_1-median. *Comput. Stat.* **27**, 393–410.

Garcia, A., Torres. J. L., Prieto, E. and De Francisco, A. (1998) Fitting wind speed distributions: A case study. *Sol. Energ.* **62**, 139–144.

Gascón, E., Lavers, D., Hamill, T. M., Richardson, D. S., Ben Bouallègue, Z., Leutbecher, M. and Pappenberger, F. (2019) Statistical post-processing of dual-resolution ensemble precipitation forecasts across Europe. *Manuscript* Available at: `https://www.esrl.noaa.gov/psd/people/tom.hamill/Calibration_DualResolution_FINALdraft.pdf` [Accessed on 16 June 2019]

Gebhardt, C., Theis, S. E., Paulat, M. and Ben Bouallègue, Z. (2011) Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.* **100**, 168–177.

Gneiting, T. (2011) Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106**, 746–762.

Gneiting, T. (2014). Calibration of medium-range weather forecasts. *ECMWF Technical Memorandum* No. 719. Available at: `http://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf` [Accessed on 16 June 2019]

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Series B Stat. Methodol.* **69**, 243–268.

Gneiting, T. and Raftery, A. E. (2005) Weather forecasting with ensemble methods. *Science* **310**, 248–249.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction and estimation. *J. Amer. Statist. Assoc.* **102**, 359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**, 1098–1118.

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Stat.* **29**, 411–422.

Gneiting, T. and Ranjan, R. (2013) Combining predictive distributions. *Electron. J. Stat.* **7**, 1747–1782.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008) Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds (with discussion and rejoinder). *Test* **17**, 211–264.

Good, I. J. (1952) Rational decisions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **14**, 107–114.

Grell, G. A., Dudhia, J. and Stauffer, D. R. (1995) A description of the fifth-generation Penn state/NCAR mesoscale model (MM5). *Technical Note* NCAR/TN-398+STR. National Center for Atmospheric Research, Boulder. Available at: `http://www2.mmm.ucar.edu/mm5/documents/mm5-desc-doc.html` [Accessed on 16 June 2019]

Hamdi, R., Degrauwe, D., Duerinckx, A., Cedilnik, J., Costa, V., Dalkilic, T., Essaouini, K., Jerczynki, M., Kocaman, F., Kullmann, L., Mahfouf, J.-F., Meier, F., Sassi, M., Schneider, S., Váňa, F. and Termonia, P. (2014) Evaluating the performance of SURFEXv5 as a new land surface scheme for the ALADINcy36 and ALARO-0 models. *Geosci. Model Dev.* **7**, 23–39.

Hamill, T. M. (2007) Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging." *Mon. Weather Rev.* **135**, 4226–4230.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y. and Lapenta, W. (2013) NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc.* **94**, 1553–1565.

Hamill, T., Engle, E., Myrick, D., Peroutka, M., Finan, C. and Scheuerer, M. (2017) The US national blend of models statistical post-processing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.* **145**, 3441–3463.

Hamill, T. M., Hagedorn, R. and Whitaker J. S. (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Weather Rev.* **136**, 2620–2632.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed., Springer, Berlin.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: a K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108.

Hemri, S., Fundel, M. and Zappa, M. (2013) Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resour. Res.* **49**, 6744–6755.

Hemri, S., Haiden, T. and Pappenberger, F. (2016) Discrete post-processing of total cloud cover ensemble forecasts. *Mon. Weather Rev.* **144**, 2565–2577.

Hemri, S., Lisniak, D. and Klein, B. (2014) Ermittlung probabilistischer Abflussvorhersagen unter Berücksichtigung zensierter Daten. *HyWa* **58**, 84–94.

Hemri, S., Lisniak, D. and Klein, B. (2015) Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resour. Res.* **51**, 7436–7451.

Hemri, S. and Klein, B. (2017) Analog based post-processing of navigation-related hydrological ensemble forecasts. *Water Resour. Res.* **53**, 9059–9077.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.* **41**, 9197–9205.

Hodyss, D., Satterfield, E., McLay, J., Hamill, T. M. and Scheuerer, M. (2016) Inaccuracies with multi-model post-processing methods involving weighted, regression-corrected forecasts. *Mon. Weather Rev.* **144**, 1649–1668.

Holman, B. P., Lazarus, S. M. and Splitt M. E. (2018) Statistically and dynamically downscaled, calibrated, probabilistic 10-m wind vector forecasts using ensemble model output statistics. *Mon. Wea. Rev.* **146** (2018), 2859–2880.

Horányi, A., Kertész, S., Kullmann, L. and Radnóti, G. (2006) The ARPEGE/ALADIN mesoscale numerical modeling system and its application at the Hungarian Meteorological Service. *Időjárás* **110**, 203–227.

Horányi, A., Mile, M. and Szűcs, M. (2011) Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary. *Tellus A* **63**, 642–651.

Iversen, T., Deckmin, A., Santos, C., Sattler, K., Bremnes, J. B., Feddersen, H. and Frogner, I.-L. (2011) Evaluation of 'GLAMEPS' – a proposed multimodel EPS for short range forecasting. *Tellus A* **63**, 513–530.

Jordan, A., Krüger, F. and Lerch, S. (2017) Evaluating probabilistic forecasts with the R package scoringRules. *arXiv* 1709.04743. Available at: `https://arxiv.org/abs/1709.04743` [Accessed on 16 June 2019].

Junk, C., Delle Monache, L. and Alessandrini, S. (2015) Analog-based ensemble model output statistics. *Mon. Weather Rev.* **143**, 2909–2917.

Justus, C. G., Hargraves, W. R., Mikhail, A. and Graber, D. (1978) Methods for estimating wind speed frequency distributions. *J. Appl. Meteor.* **17**, 350–353.

Kain, J. S. and Fritsch, J. M. (1990) A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.* **47**, 2784–2802.

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. and Grimit, E. (2011) Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Weather Rev.* **139**, 2630–2649.

Knüppel, M. (2015) Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *J. Bus. Econ. Stat.* **33**, 270–281.

Krzysztofowicz, R. (1999) Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.* **35**, 2739–2750.

Lindström, G., Johansson, B., Persson, M., Gardelin, M. and Bergström, S. (1997) Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **201**, 272–288.

Lee, G and Scott, C. (2012) EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Comput. Stat. Data Anal.* **56**, 2816–2829.

Leith, C. E. (1974) Theoretical skill of Monte-Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418.

Lerch, S. and **Baran, S.** (2017) Similarity-based semi-local estimation of EMOS models. *J. R. Stat. Soc. Ser. C Appl. Statist.* **66**, 29–51.

Lerch, S. and Thorarinsdottir, T. L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65**, 21206.

Leutbecher, M. and Palmer, T. N. (2008) Ensemble forecasting. *J. Comp. Phys.* **227**, 3515–3539.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions.* Wiley, New York.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Möller, A. and Groß, J. (2016) Probabilistic temperature forecasting based on an ensemble AR modification. *Q. J. R. Meteorol. Soc.* **142**, 1385–1394.

Möller, A., Lenkoski, A. and Thorarinsdottir, T. L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.* **139**, 982–991.

Möller, A., Lenkoski, A., Thorarinsdottir, T. L. and Gneiting, T. (2015) Spatially adaptive, Bayesian estimation for probabilistic temperature forecasts. *arXiv*:1507.05066. Available at: `https://arxiv.org/abs/1507.05066` [Accessed on 16 June 2019]

Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T. (2011) Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: Main achievements and open challenges. *Tellus A* **63**, 605–624.

Murphy, A. H. (1973) Hedging and skill scores for probability forecasts. *J. Appl. Meteorol.* **12**, 215–223.

National Weather Service (1998) *Automated Surface Observing System (ASOS) User's Guide.* Available at: `https://www.weather.gov/media/asos/aum-toc.pdf` [Accessed on 16 June 2019]

Nelder, J. A. and Mead, R. (1965) A simplex algorithm for function minimization. *Comput. J.* **7**, 308–313.

Noilhan, J. and Planton, S. (1989) A simple parameterization of land surface processes for meteorological models. *Mon. Weather Rev.* **117**, 536–549.

Park, Y.-Y., Buizza, R. and Leutbecher, M. (2008) TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.* **134**, 2029–2050.

Pinson, P. (2012) Adaptive calibration of (u, v)-wind ensemble forecasts. *Q. J. R. Meteorol. Soc.* **138**, 1273–1284.

Pinson, P. and Girard, R. (2012) Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy* **96**, 12–20.

Pinson, P and Hagedorn, R. (2012) Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorol. Appl.* **19**, 484–500.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. T. (2007) *Numerical Recipes 3rd Edition: The Art of Scientific Computing.* Cambridge University Press, Cambridge.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174.

Raftery, A. E, Kárný, M. and Ettler, P. (2010) Online prediction under model uncertainty via Dynamic Model Averaging: application to a cold rolling mill. *Technometrics* **52**, 52–66.

R Core Team (2019) *R: A Language and Environment for Statistical Computing.* Available at: `https://www.R-project.org/` [Accessed on 16 June 2019]

Richardson, D., Hemri, S., Bogner, K., Gneiting, T., Haiden. T., Pappenberger, F. and Scheuerer, M. (2015) Calibration of ECMWF forecasts. *ECMWF Newsletter* **142**, 12–16.

Sass, B. H. (2002) A research version of the STRACO cloud scheme. *DMI Tech. Rep.* 02-10. Danish Meteorological Institute, Copenhagen, Denmark, 25 pp. Available at: `https://www.researchgate.net/publication/239846781_A_research_version_of_the_STRACO_cloud_scheme` [Accessed on 16 June 2019]

Schefzik, R. (2016a) A similarity-based implementation of the Schaake shuffle. *Mon. Weather Rev.* **144**, 1909–1921.

Schefzik, R. (2016b) Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Q. J. R. Meteorol. Soc.* **142**, 2463–2477.

Schefzik, R., Thorarinsdottir T. L. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28**, 616–640.

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* **140**, 1086–1096.

Scheuerer, M. and Büermann, L. (2014) Spatially adaptive post-processing of ensemble forecasts for temperature. *J. R. Stat. Soc. Ser. C Appl. Statist.* **63**, 405–422.

Scheuerer, M. and Hamill, T. M. (2015) Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **143**, 4578–4596.

Scheuerer, M. and Möller, D. (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.* **9**, 1328–1349.

Schuhen, N., Thorarinsdottir, T. L. and Gneiting, T. (2012) Ensemble model output statistics for wind vectors. *Mon. Weather Rev.* **140**, 3204–3219.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Am. Stat. Assoc.* **105**, 25–37.

Sloughter, J. M., Gneiting, T and Raftery, A. E. (2013) Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Mon. Weather Rev.* **141**, 2107–2119.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Weather Rev.* **135**, 3209–3220.

Stathopoulou, M. I., Cartalis, C., Keramitsoglou, I. and Santamouris, M. (2005) Thermal remote sensing of Thom's discomfort index (DI): comparison with in-situ measurements. *Proc. SPIE 5983, Remote Sensing for Environmental Monitoring, GIS Applications, and Geology.* 59830K (29 October 2005); doi:10.1117/12.627541.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J. H., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H. A., Wilson, L. and Yamaguchi, M. (2016) The TIGGE project and its achievements. *B. Am. Meteorol. Soc.* **97**, 49–67.

Thorarinsdottir, T. L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *J. R. Stat. Soc. Ser. A Stat. Soc.* **173**, 371–388.

Todini, E. (2008) A model conditional processor to assess predictive uncertainty in flood forecasting. *Int. Jour. of River Basin Manag.* **6**, 123–137.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.* **125**, 3297–3319.

Vannitsem, S., Wilks, D. S., Messner, J. W. (eds.) (2018) *Statistical Postprocessing of Ensemble Forecasts.* Elsevier, Amsterdam.

van Straaten, C., Whan, K. and Schmeits, M. (2018) Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *J. Hydrometeorol.* **19**, 1815–1833.

Vardi, Y. and Zhang, C. H. (2000) The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97**, 1423–1426.

Wilks, D. S. (2006) Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorol. Appl.* **13**, 243–256.

Wilks, D. S. (2009) Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorol. Appl.* **16**, 361–368.

Wilks, D. S. (2011) *Statistical Methods in the Atmospheric Sciences.* 3rd ed., Elsevier, Amsterdam.

Wilks, D. S. (2018) Univariate ensemble forecasting. In Vannitsem, S., Wilks, D. S., Messner, J. W. (eds.), *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, pp. 49–89.

Williams, R. M., Ferro, C. A. T. and Kwasniok F (2014) A comparison of ensemble post-processing methods for extreme events. *Q. J. R. Meteorol. Soc.* **140**, 1112–1120.

Yang, X., Sharma, S., Siddique, R., Greybush, S. J. and Mejia, A. (2017) Postprocessing of GEFS precipitation ensemble reforecasts over the US Mid-Atlantic region. *Mon. Weather Rev.* **145**, 1641–1658.

Yuen, R. A., **Baran, S.**, Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M., Thorarinsdottir, T. L. (2018) *R package ensembleMOS, Version 0.8.2: Ensemble Model Output Statistics.* Available at: `https://cran.r-project.org/package=ensembleMOS` [Accessed on 16 June 2019]